ZC3H4 restricts non-coding transcription in human cells

ZC3H4 restricts non-coding transcription in human cells

Chris Estell[1], Lee Davidson[1], Pieter C. Steketee[2], Adam Monier[1] and Steven West[1]*

[1]The Living Systems Institute, University of Exeter, Stocker Rd, Exeter, EX4 4QD

[2] The Roslin Institute, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, EH25 9RG

*Lead contact: s.west@exeter.ac.uk; telephone 0044(0)1392727458

Keywords: ZC3H4; Transcription termination; enhancer; non-coding RNA; super-enhancer; exosome; antisense; RNA polymerase II; Integrator

SUMMARY

The human genome encodes thousands of non-coding RNAs. Many of these terminate early and are then rapidly degraded, but how their transcription is restricted is poorly understood. In a screen for protein-coding gene transcriptional termination factors, we identified ZC3H4. Its depletion causes upregulation and extension of hundreds of unstable transcripts, particularly antisense RNAs and those transcribed from so-called super-enhancers. These loci are occupied by ZC3H4, suggesting that it directly functions in their transcription. Consistently, engineered tethering of ZC3H4 to reporter RNA promotes its degradation by the exosome. ZC3H4 is predominantly metazoan - interesting when considering its impact on enhancer RNAs that are less prominent in single-celled organisms. Finally, ZC3H4 loss causes a substantial reduction in cell proliferation, highlighting its overall importance. In summary, we identify ZC3H4 as playing an important role in restricting non-coding transcription in multi-cellular organisms.

INTRODUCTION

Most of the human genome can be transcribed by RNA polymerase II (Pol II). Among these transcripts are thousands of long non-coding RNAs, broadly classified as greater than ~200 nucleotides in length (Kopp and Mendell, 2018). They share some structural features with coding transcripts, but most of them are rapidly degraded by the exosome (Davidson et al., 2019; Preker et al., 2008; Schlackow et al., 2017). Their degradation is coincident with or shortly after transcriptional termination, which often occurs within a few kilobases (kb). The mechanisms for terminating non-coding transcription are poorly understood, especially by comparison with those operating at protein-coding genes.

Termination of protein-coding transcription is coupled to 3' end processing of pre-mRNA via cleavage at the polyadenylation signal (PAS) (Eaton and West, 2020). A PAS consists of an AAUAAA hexamer followed by a U/GU-rich region (Proudfoot, 2011). After assembly of a multi-protein processing complex, CPSF73 cleaves the nascent RNA and the Pol II-associated product is degraded 5'→3' by XRN2 to promote termination (Eaton et al., 2018; Eaton et al., 2020; Fong et al., 2015). The Pol II elongation complex is modified as it crosses the PAS, which facilitates its termination by XRN2 (Cortazar et al., 2019; Eaton et al., 2020). Depletion of XRN2 or CPSF73 causes read-through downstream of some long non-coding genes (Eaton et al., 2020). However, a substantial fraction of non-coding transcription is less sensitive to their depletion suggesting the use of alternative mechanisms.

The Integrator complex aids termination of many non-coding transcripts, with the archetypal example being snRNAs (Baillat et al., 2005; Davidson et al., 2020; O'Reilly et al., 2014). Integrator is also implicated in the termination of promoter upstream transcripts (PROMPTs) and enhancer RNAs (eRNAs) (Beckedorff et al., 2020; Lai et al., 2015; Nojima et al., 2018). The mechanism is analogous to that at protein-coding genes, driven by endonucleolytic cleavage by INTS11. However, INTS11 activity does not precede XRN2 degradation at snRNA genes (Eaton et al., 2018). Moreover, while CPSF73 is indispensable for termination at protein-coding genes, there is evidence of redundant pathways at snRNA loci (Davidson et al., 2020). Indeed, CPSF and the cap binding complex-associated factor, ARS2, are both implicated in the termination of promoter-proximal transcription (Iasillo et al., 2017; Nojima et al., 2015).

A variety of processes attenuate transcription at protein-coding genes (Kamieniarz-Gdula and Proudfoot, 2019). Frequently, this is via premature cleavage and polyadenylation (PCPA) that can be controlled by U1 snRNA, CDK12, SCAF4/8 or PCF11 (Dubbury et al., 2018; Gregersen et al., 2019; Kaida et al., 2010; Kamieniarz-Gdula et al., 2019). PCPA is

common on many genes since acute depletion of the exosome stabilises its predicted products in otherwise unmodified cells (Chiu et al., 2018; Davidson et al., 2019). Integrator activity also attenuates transcription at hundreds of protein-coding genes.

A less-studied termination pathway at some intragenic non-coding regions is controlled by WDR82 and its associated factors (Austenaa et al., 2015). In mammals, WDR82 forms at least two complexes: one with the SETD1 histone methyl-transferase and another composed of protein-phosphatase 1 and its nuclear targeting subunit PNUTS (Lee et al., 2010; van Nuland et al., 2013). A version of the latter promotes transcriptional termination in trypanosomes (Kieft et al., 2020) and the budding yeast homologue of WDR82, Swd2, forms part of the APT (associated with Pta1) termination complex (Nedea et al., 2003). In murine cells, depletion of either WDR82, PNUTS or SET1 causes non-coding transcriptional termination defects (Austenaa et al., 2015). Notably, PNUTS/PP1 is implicated in the canonical termination pathway at protein-coding genes where its dephosphorylation of SPT5 causes deceleration of Pol II beyond the PAS (Cortazar et al., 2019; Eaton et al., 2020).

Here, we performed a proteomic screen for new termination factors by searching for proteins that bind to Pol II complexes in a manner that depends on PAS recognition by CPSF30. This uncovered ZC3H4, a metazoan zinc finger-containing factor without a characterised function in transcription. Because of the nature of our screen, we anticipated a role for ZC3H4 in 3' end formation; however, its effects on this process are mild and apply to a small number of genes. Its main function is to restrict non-coding transcription, especially of PROMPT and eRNA transcripts, which are extended by hundreds of kb when ZC3H4 is depleted. ZC3H4 interacts with WDR82, the depletion of which causes similar defects. Tethered function assays show that ZC3H4 recruitment is sufficient to restrict transcription and cause RNA degradation by the exosome. In sum, we reveal ZC3H4 as a hitherto unknown terminator of promoter-proximal transcription with particular relevance at non-coding loci.

RESULTS

The effect of CPSF30 depletion on the Pol II-proximal proteome

The first step of PAS recognition involves the binding of CPSF30 to the AAUAAA signal (Chan et al., 2014; Clerici et al., 2018; Sun et al., 2018). We reasoned that elimination of CPSF30 would impede PAS-dependent remodelling of elongation complexes and cause the retention or exclusion of potentially undiscovered transcriptional termination

factors. We used CRISPR/Cas9 genome editing to tag *CPSF30* with a mini auxin-inducible degron (mAID) (Figure 1A). The integration was performed in HCT116 cells where we had previously introduced the plant F-box gene, *TIR1*, required for the AID system to work (Eaton et al., 2018; Natsume et al., 2016). CPSF30-mAID is eliminated by 3 hours of indol-3-acetic acid (auxin/IAA) treatment (Figure 1B). This results in profound and general transcriptional read-through downstream of protein-coding genes (Figure 1C and Figure 1-figure supplement 1A) demonstrating widespread impairment of PAS function.

To identify Pol II interactions sensitive to CPSF30, we further modified *CPSF30-mAID* cells to homozygously tag the largest subunit of Pol II, Rpb1, with mini(m)-Turbo (Figure 1D and Figure 1- figure supplement 1B). mTurbo is an engineered ligase that biotinylates proximal proteins when cells are exposed to biotin (Branon et al., 2018). This occurs within minutes of biotin addition to culture media, which is advantageous for analysing dynamic proteins such as Pol II. We chose this approach rather than immunoprecipitation (IP) because it allows isolation of weak/transient interactions (potentially disrupted during conventional IP) and may identify relevant proximal proteins that do not interact with Pol II directly. Importantly, CPSF30-mAID depletion still induced strong read-through in this cell line (Figure 1 - figure supplement 1C).

CPSF30-mAID:RPB1-mTurbo cells were exposed to biotin before western blotting with streptavidin horseradish peroxidase (HRP). This revealed multiple bands with a prominent one corresponding in size to Rpb1-mTurbo and indicating the biotinylation of its proximal proteome (Figure 1E). A small number of endogenously biotinylated factors were observed in the absence of biotin. Biotin-exposed samples were subject to tandem mass tagging (TMT) with mass spectrometry. We focused on proteins with reduced abundance after auxin treatment (Supplementary File 1). The factor most depleted was CPSF30, confirming that its auxin-dependent depletion is reflected in the data (Figure 1F). As expected, Rpb1 was the most abundant factor in all samples consistent with its self-biotinylation seen by western blotting. After CPSF30, the most depleted factors were Fip1, CPSF100 and WDR33, which are in the CPSF complex. Otherwise, surprisingly few proteins showed reduced signal following auxin treatment. This implies that the major effect of CPSF30 depletion on the Pol II-proximal interactome is to prevent the recruitment/assembly of the CPSF complex.

ZC3H4 is a candidate transcription termination factor that is metazoan-enriched

Two poorly characterised factors, ZC3H4 and ZC3H6, were the next most depleted. They contain CCCH zinc finger motifs flanked by intrinsically disordered regions (Figure 1 –

figure supplement 2A). Their potential relationship to canonical 3' end formation factors is suggested via known/predicted protein-protein interactions that are collated by the STRING database (Jensen et al., 2009) (Figure 1 – figure supplement 2B). ZC3H4 is also co-regulated with mRNA processing factors suggesting a role in RNA biogenesis (Figure 1 – figure supplement 2C; (Kustatscher et al., 2019)). Although little is reported on ZC3H4, two independent studies uncovered it as an interaction partner of WDR82 using Mass Spectrometry (Lee et al., 2010; van Nuland et al., 2013). WDR82 plays a key role in transcriptional termination in yeast, trypanosomes and mice (Austenaa et al., 2015; Kieft et al., 2020; Nedea et al., 2003). To verify this interaction, we tagged ZC3H4 with GFP and performed a "GFP trap" whereby ZC3H4-GFP is captured from whole cell lysates using GFP nanobody-coupled beads (Figure 1 – figure supplement 2D). WDR82 robustly co-precipitated with ZC3H4-GFP confirming them as interacting partners. Although WDR82 is conserved between human and budding yeast our phylogenetic analysis suggested that ZC3H4 and ZC3H6 are largely restricted to metazoans and are paralogues (Figure 1 - figure supplement 3A and B).

ZC3H4 restricts non-coding transcription events

To assess any function of ZC3H4 and/or ZC3H6 in RNA biogenesis we depleted either or both from HCT116 cells using RNA interference (RNAi) (Figure 2 figure supplement 1A), then deep sequenced nuclear transcripts. Comparison of these datasets shows that ZC3H4 loss has a more noticeable impact than ZC3H6 depletion (Figure 2 – figure supplement 1B). Specifically, ZC3H6 depleted samples are more similar to control than those deriving from ZC3H4 loss and ZC3H4/ZC3H6 co-depletion resembles a knock-down of just ZC3H4. This was also evident from closer inspection of the data (Figure 2 – figure supplement 1C), supporting the phylogenetic prediction of their separate functions. Accordingly, subsequent analyses focus on ZC3H4.

Due to its links with CPSF30 and WDR82, we anticipated that ZC3H4 might affect transcriptional termination. We first checked protein-coding genes and found a small number with longer read-through beyond the PAS when ZC3H4 is depleted (Figure 2A). However, broader analysis suggests that this is not widespread and far fewer genes exhibit increased read-through following ZC3H4 loss compared to when CPSF30 is absent (Figure 2B and Figure 2 – figure supplement 2A-D). Interestingly, the metagene in Figure 2B revealed slightly more signal antisense of promoters when ZC3H4 is depleted. This indicates an effect on non-coding RNA, which is interesting in light of a previously described function for WDR82 in restricting intragenic transcription (Austenaa et al., 2015). These

PROMPT transcripts are normally rapidly degraded 3'→5' by the exosome (Preker et al., 2008). Figure 2C shows an example PROMPT, upstream of *MYC*, which is undetectable in control siRNA treated cells, but abundant following ZC3H4 depletion. Loss of ZC3H4 also leads to the extension of this transcript by more than 100 kilobases. This is made clearer by comparing the loss of ZC3H4 to AID-mediated depletion of the catalytic exosome (DIS3) (Davidson et al., 2019). DIS3 depletion stabilises the usual extent of PROMPT RNA, which is much shorter than when ZC3H4 is absent. Importantly, meta-analysis reveals similar effects at many other PROMPTs (Figure 2D). These data strongly suggest that PROMPT transcripts are stabilised and extended in the absence of ZC3H4, presumably because its normal function restricts their transcription.

The finding that PROMPTs are affected by ZC3H4 suggested a role in the transcription/metabolism of antisense/non-coding RNAs. We therefore extended our search for potential ZC3H4 regulated transcription to enhancer regions since they also produce short RNAs that are degraded by the exosome (Andersson et al., 2014). eRNAs can be found in isolation and in clusters called super-enhancers (SEs) (Pott and Lieb, 2015). SEs are thought to be important for controlling key developmental genes with strong relevance to disease (Hnisz et al., 2013). ZC3H4 depletion has a clear effect over SE regions exemplified by the *MYC* SE where upregulation and extension of eRNAs is obvious (Figure 2E). Acute depletion of DIS3 illustrates the normally restricted range of individual eRNAs within the cluster. This effect is general for other SEs as demonstrated by the metaplots in Figure 2F. We also checked the effect of CPSF30 depletion on example PROMPT and SE transcription, which are very modest and consistent with the lack of antisense effect seen by metagene in Figure 1C (Figure 2 – figure supplement 2E). Consistently, PROMPTs susceptible to ZC3H4 were not enriched in PASs compared to those unaffected by it and harbour a slightly lower density (Figure 2 – figure supplement 2F). Overall, these data strongly suggest that ZC3H4 is important for regulating transcription across many PROMPTs and SEs.

Comparison of ZC3H4 and Integrator effects

ZC3H4 has some functions in common with the Integrator complex. This is a metazoan-specific assembly with regulatory functions at non-coding loci (Lai et al., 2015; Mendoza-Figueroa et al., 2020; Nojima et al., 2018). We previously sequenced chromatin-associated RNA derived from HCT116 cells RNAi depleted of the Integrator backbone component INTS1 (Davidson et al., 2020). Chromatin-associated RNA is purified via urea/detergent extraction and is enriched in nascent RNAs (Wuarin and Schibler, 1994). Metagene analysis of this data at protein-coding genes shows a mild effect of Integrator

depletion over PROMPT regions (Figure 3A). It also reveals an accumulation of promoter-proximal RNAs in the coding direction consistent with a recent report on its function as an attenuator of protein-coding transcription (Lykke-Andersen et al., 2020). Because of this function, Integrator depletion can lead to increased expression of a subset of mRNAs (Elrod et al., 2019; Lykke-Andersen et al., 2020; Tatomer et al., 2019). *HAP1* is an example of a gene where this is seen (Figure 3B). Similarly, we saw evidence for increased mRNA expression on some genes when ZC3H4 was depleted (Figure 3C). Interestingly these two genes are selectively effected by Integrator or ZC3H4 respectively and additional examples of this are shown in Figure 3 – figure supplement 1A. Bioinformatic analysis revealed around 1000 genes affected by INTS1 or ZC3H4 depletion with little overlap between the two conditions (Figure 3D, Supplementary File 4). Indeed, analysis of recently published metabolically labelled RNA-seq data from HeLa cells depleted of the catalytic Integrator subunit or ZC3H4 reveals several hundred upregulated mRNAs - also with minimal overlap (Austenaa et al., 2021; Lykke-Andersen et al., 2020) (Figure 3 – figure supplement 1B). When searching for characteristics of these targets in our HCT116 data, we found that transcripts upregulated following either ZC3H4 or INTS1 loss are normally expressed at lower levels than unaffected genes (Figure 3E). This is consistent with the idea that they are subject to repression by these two factors under these experimental conditions.

The most prominent effects of ZC3H4 were observed at PROMPT and SE regions where, again, Integrator is implicated (Lai et al., 2015; Nojima et al., 2018). Where ZC3H4 effects are evident over PROMPT regions, they are generally more substantial than those seen after Integrator loss, exemplified by the *ITPRID2* PROMPT in Figure 3F and via meta-analyses (Figure 3 – figure supplement 1C and D). At SEs, ZC3H4 depletion generally results in a greater stabilisation and elongation of eRNA, compared to INTS1 knock-down, exemplified at the *MSRB3* SE (Figure 3G). Meta-analysis confirms less effect of INTS1 depletion at SEs versus the impact of ZC3H4 (compare Figures 3H and 2F). We note that these INTS1 data are on chromatin-associated RNA whereas ZC3H4 images are obtained from nuclear RNA. However, as chromatin-associated RNA is more enriched in nascent transcripts this would be expected to capture more extended non-coding transcription and not less as is the case here. Moreover, previously published analyses of Integrator effects on transcription do not report the long extended non-coding (PROMPT/eRNA) transcripts that we observe when ZC3H4 is depleted (Beckedorff et al., 2020; Lykke-Andersen et al., 2020).

Rapid ZC3H4 depletion and re-expression confirms the functions found by RNA-seq

ZC3H4 RNAi suggests its widespread involvement in non-coding RNA synthesis and the regulation of a subset of protein-coding transcripts. However, RNAi depletion was performed using a 72 hr protocol and might result in indirect or compensatory effects. To assess whether these effects are a more direct consequence of ZC3H4 loss, we engineered HCT116 cells for its rapid and inducible depletion. CRISPR/Cas9 was used to tag *ZC3H4* with an *E.coli* derived DHFR degron preceded by 3xHA epitopes (Figure 4A; (Sheridan and Bentley, 2016)). In this system, cells are maintained in trimethoprim (TMP) to stabilise the degron, removal of which causes protein depletion. Western blotting demonstrates homozygous tagging of *ZC3H4* and that ZC3H4-DHFR is depleted following TMP removal (Figure 4B). Depletion was complete after overnight growth without TMP but substantial protein loss was already observed after 4 hrs allowing us to assess the consequences of more rapid ZC3H4 depletion.

TMP-mediated depletion can also be reversed by its re-administration facilitating a test of whether ZC3H4 effects are reversed by its reappearance. The western blot in Figure 4C illustrates this by showing that TMP withdrawal depletes ZC3H4-DHFR, which re-appears following 4 hrs TMP addition. To ask whether ZC3H4 effects are an immediate consequence of its loss and if they are reversed following its re-appearance, RNA was isolated from the three conditions shown in the western blot. This was analysed by qRT-PCR to assess the levels of extended PROMPT (*HMGA2, ITPRID2*) and SE (*MSRB3, DLGAP1*) RNAs (Figure 4D). All were increased following ZC3H4 loss suggesting that the effects that we observed by RNAi are not due to compensatory pathways. Although 4 hr TMP re-administration does not restore ZC3H4 to full levels, it was sufficient to reverse the effects of its depletion at all tested amplicons. The timescale over which the effect can be reversed suggests that transcripts induced by ZC3H4 loss remain relatively unstable. Rapid ZC3H4 depletion also confirmed the prediction, from our RNA-seq, that the extended PROMPT transcripts result from the aberrant transcription of these loci (Figure 4 – figure supplement 1A and B).

Another key observation from our nuclear RNA-seq was the potential for ZC3H4 to restrict the levels of a subset of protein-coding transcripts. The long-term nature of RNAi and its detection via nuclear RNA-seq means that it could be an indirect or post-transcriptional effect, respectively. To test whether mRNA upregulation is an immediate and transcriptional response to ZC3H4 loss, we isolated chromatin-associated RNA from *ZC3H4-DHFR* cells grown with or without TMP for 4 hrs. To additionally confirm their specificity to ZC3H4 (vs Integrator), we also depleted the catalytic Integrator subunit utilising our previously engineered cell line in which INTS11 is tagged with a small molecule assisted shut-off module (Chung et al., 2015; Davidson et al., 2020). qRT-PCR was used to detect

three transcripts (NWD1, ENO3 and PJVK) that were upregulated by ZC3H4 loss but not Integrator depletion. Spliced versions of all three were increased after 4 hrs of ZC3H4 depletion, but unaffected by loss of the catalytic Integrator subunit INTS11 (Figure 4E). The effectiveness of INTS11 depletion is illustrated by the substantial increase in U1 snRNA read-through RNA in its absence. This demonstrates that some mRNAs are immediately and selectively upregulated following ZC3H4 loss.

We next asked whether the ZC3H4 interactor, WDR82, impacts the levels of PROMPT and SE transcripts. Accordingly, *ZC3H4-DHFR* cells were treated with control or WDR82-specific siRNAs (Figure 4F). We also co-depleted ZC3H4 and WDR82 by removing TMP from cells first transfected with WDR82 siRNAs. WDR82 depletion enhanced the level of all tested transcripts suggesting that it functionally overlaps with ZC3H4 (Figure 4G). There was no synergistic effect of their co-depletion implying that WDR82 and ZC3H4 do not act redundantly at the tested loci. WDR82 is found in complexes containing protein phosphatase 1 (PP1) and the SETD1A/B methyl transferases (Lee et al., 2010; van Nuland et al., 2013). We found that the former but not the latter is implicated in the stability of the non-coding transcripts selected for this experiment (Figure 4 – figure supplement 1C-E).

ZC3H4 occupies a broad region at a subset of promoters

We have demonstrated that depletion of ZC3H4 causes widespread defects in non-coding transcription and supresses a subset of protein-coding RNAs. As these effects are seen following rapid ZC3H4 depletion, we hypothesised that they may be directly mediated by its recruitment to relevant loci. Consistently, its capture in our mTurbo experiment supports its proximity to chromatin and the presence of CCCH zinc finger domains predict nucleic acid binding capability. Therefore, its genomic occupancy was globally investigated by performing ZC3H4 chromatin immunoprecipitation and sequencing (ChIP-seq) alongside that of Pol II.

ZC3H4 occupies genes with binding broadly resembling that of Pol II and showing the greatest enrichment over promoter regions (Figure 5A). However, many genes that are occupied by Pol II do not recruit ZC3H4 (Figure 5B). This might result from low affinity of the ZC3H4 antibody or that its recruitment to chromatin is bridged since ZC3H4 also directly crosslinks to RNA in cells (Figure 5 – figure supplement 1A). However, its differential gene occupancy is consistent with the selective effects of its depletion. Interestingly, ZC3H4 occupies a broader promoter region than Pol II suggesting that its function is not restricted to the precise transcriptional start site. The width of this peak often corresponds to the normal extent of PROMPT and eRNA transcription, which is elongated in its absence. *RPL13* is shown as an example of recruitment of ZC3H4 upstream of the promoter, where its loss

causes stabilisation and extension of the antisense transcript (Figure 5C). ZC3H4 is also strongly recruited to SEs consistent with the RNA effects observed on them following its loss (Figure 5D). This is exemplified by the *MSRB3* region and generalised by metaplots in figure 5E. Although our analyses of eRNA and PROMPTs were guided by our RNA-seq findings, an unbiased search for peaks of ZC3H4 and Pol II signal confirmed proportionally greater ZC3H4 occupancy at distal intergenic regions (encompassing SEs) (Figure 5F).

Overall, the HCT116 ChIP-seq demonstrates direct recruitment of ZC3H4 to potential targets. One mentioned caveat is the low ChIP efficiency of the ZC3H4 antibody; however, a ZC3H4 ChIP-seq experiment was recently made available by the ENCODE consortium (Partridge et al., 2020). This used a flag-tagged construct and was performed in HEPG2 cells allowing a comparison of our data to that obtained with a high-affinity antibody and in different cells. Consistent with our findings, flag-ZC3H4 occupies a subset of Pol II-bound regions and shows broader distribution than Pol II around promoters (Figure 5G). Although HEPG2 cells express fewer SEs than HCT116 cells, the transcribed *DLGAP1* example confirms its occupancy of these regions in both cell types (Figure 5 – figure supplement 1B). In contrast, the *MYC* SE is only expressed in HCT116 cells and is not occupied by ZC3H4 in HEPG2 cells. In further agreement with our data, bioinformatics assignment of flag-ZC3H4 binding sites yielded “promoter and enhancer-like” as the most enriched terms (Partridge et al., 2020).

Engineered recruitment of ZC3H4 suppresses transcription

The consequences of ZC3H4 recruitment to targets are predicted to be their early termination and subsequent degradation by the exosome, based on the known fate of PROMPTs and eRNAs. To test whether ZC3H4 recruitment can promote these effects, we established a tethered function assay. ZC3H4 was tagged with bacteriophage MS2 coat protein to engineer its recruitment to a reporter containing MS2 hairpin binding sites (MS2hp-IRES-GFP; Figure 6A). Importantly, RNA from this reporter is unaffected by endogenous ZC3H4 (Figure 6 – figure supplement 1A). HCT116 cells were transfected with either of these three constructs together with MS2hp-IRES-GFP and reporter expression assayed by qRT-PCR. Compared to the two controls, tethered ZC3H4-MS2 significantly reduced reporter RNA expression (Figure 6B). ZC3H4-MS2 expression does not affect the same reporter lacking MS2 hairpins (Figure 6 – figure supplement 1B). This directly demonstrates that ZC3H4 recruitment is sufficient to negatively regulate RNA expression, mirroring the upregulation of its endogenous targets seen when it is depleted.

PROMPTs and eRNAs are degraded on chromatin and we wanted to test whether ZC3H4-MS2 exerted its effect on these nascent RNAs. The reporter experiments above are on total RNA so whether ZC3H4-MS2 exerted its effect at the gene (plasmid) or following its release was uncertain. Therefore, we purified chromatin-associated RNA using a salt and urea-based extraction (Wuarin and Schibler, 1994). As mentioned previously, this fractionation enriches nascent endogenous RNAs. However, nascent RNAs associated with transfected plasmids also co-purify within this fraction (Dye et al., 2006). Accordingly, cells were transfected with MS2hp-IRES-GFP and either ZC3H4-MS2 or MS2-GFP. We included an additional primer set to detect RNA uncleaved at the bovine growth hormone (BGH) poly(A) site. Because poly(A) site cleavage is co-transcriptional, this primer set should robustly detect Pol II-associated transcripts. This amplicon and that upstream of the MS2 hairpins were reduced in this chromatin fraction, strongly suggesting that tethered ZC3H4 acts on nascent RNA (Figure 6C).

The exosome targets released PROMPT and eRNA transcripts, which could be promoted by ZC3H4. The results we present for endogenous loci are consistent with this since PROMPTs and eRNAs are upregulated and elongated when ZC3H4 is depleted. To test whether recruited ZC3H4 leads to exosome decay, we transfected MS2hp-IRES-GFP, together with either MS2-GFP or ZC3H4-MS2, into *DIS3-AID* cells that were then treated or not with auxin to eliminate the catalytic exosome. RNA upstream and downstream of the MS2 hairpins was detected by qRT-PCR and their ratio plotted (Figure 6D). Enhanced levels of upstream versus downstream amplicon were associated with transfection of ZC3H4-MS2 and is more prominent after depletion of DIS3. This is consistent with the hypothesis that recruited ZC3H4 promotes the release of RNA that is a DIS3 substrate (Figure 6E). Results presented above show that ZC3H4 functions in transcriptional regulation. ZC3H4 may also regulate the stability of its targets; however, to our knowledge, it has not been found to prominently co-purify with the exosome.

Finally, we were interested to determine the overall relevance of ZC3H4 to cell health/growth. This is made simpler by the *ZC3H4-DHFR* cell line, which allows permanent depletion of ZC3H4 by culturing cells without TMP. Accordingly, we performed colony formation assays on these cells grown in the presence or absence of TMP (Figure 6F). Loss of ZC3H4 was associated with smaller colonies, which demonstrates the importance of ZC3H4 for growth/proliferation.

DISCUSSION

We have discovered that ZC3H4 controls unproductive transcription, especially at non-coding loci. This conclusion is based on its recruitment to loci that give rise to transcripts that are stabilised and elongated when it is depleted. Moreover, tethering ZC3H4 to a heterologous reporter RNA is sufficient to promote degradation of the transcript by the exosome. We propose that ZC3H4 recruitment drives some of the early transcriptional termination that is characteristic of many non-coding RNAs, particularly PROMPT and eRNA transcripts. The function of ZC3H4 in restraining their transcription may at least partly explain why PROMPT and eRNA transcripts accumulate as short species when the exosome is depleted.

Our discovery of ZC3H4 adds to an increasing number of termination pathways. Most of these are more relevant during the initial stages of transcription, rather than the more intensively studied process that occurs at the end of protein-coding genes. This is evident from comparing the general requirement for CPSF30 at the 3' end of protein-coding genes with the more selective impact of ZC3H4 that is focused more promoter-proximally. The effects of ZC3H4 depletion are reminiscent of recent findings on the Integrator complex, which also controls the early termination of transcription (Elrod et al., 2019; Lykke-Andersen et al., 2020; Tatomer et al., 2019). Our initial comparison of transcripts sensitive to either Integrator or ZC3H4 suggests that they can act on separate RNA targets. An exciting possibility is that multiple early termination pathways may contribute to conditional gene regulation. It will be important to establish whether ZC3H4 and/or Integrator are naturally utilised to regulate transcription in this manner. Their predominance in metazoans may enable complex gene regulation, for example across cell types or during development.

ZC3H4 has been proposed as an equivalent to *Drosophila* Suppressor of Sable (Su(s)), which negatively regulates transcription via promoter-proximal termination (Brewer-Jensen et al., 2016; Kuan et al., 2004). ZC3H4 and Su(s) share little sequence homology, but they have similar structural makeup with zinc fingers flanked by largely disordered regions. Su(s) depletion stabilises selected RNAs and causes their aberrant elongation and stability, mirroring what we see globally following ZC3H4 depletion. There is no known catalytic activity for ZC3H4 or Su(s), but they are related to CPSF30 which shows endonuclease activity in *Drosophila* and *Arabidopsis* (Addepalli and Hunt, 2007; Bai and Tolias, 1996). It remains to be seen whether ZC3H4 possesses any catalytic activity or mediates its effects through interaction partners. Interestingly, IP and mass spectrometry indicates that WDR82 may be the only interacting partner of Su(s) (Brewer-Jensen et al., 2016). WDR82 has been shown to bind to Pol II phosphorylated on Serine 5 of its C-terminal domain, which may provide a means to recruit ZC3H4 to promoter-proximal regions (Lee and Skalnik, 2008).

The recruitment of ZC3H4 to promoters is consistent with our observation that promoter-proximal transcription is most affected by its absence. Since depletion of ZC3H4 causes extended transcription of its targets, it is reasonable to suppose that it normally restricts their transcription in some fashion. This might be by controlling the escape of promoter-proximally paused polymerases or by acting closer to the 3' end of its target transcripts. The fact that ZC3H4 acts somewhat selectively (e.g. not all PROMPTs and mRNAs are its targets) suggests that elements of specificity are required to explain its mechanism. Most obviously, this could be sequences within DNA or RNA, to which ZC3H4 (and Su(s)) binds via ChIP and XRNAX, respectively (see Figure 5 and Figure 5 – figure supplement 1A). While our paper was under revision, another report identified ZC3H4 as affecting the transcription of intragenic loci in mammalian cells (Austenaa et al., 2021). In agreement with our findings, non-coding transcripts were affected by ZC3H4 depletion. It was proposed to terminate some non-coding transcripts as a result of spurious/weak splicing. Similarly, Su(s) regulation of transcription was linked to the presence of a cryptic 5' splice site (Kuan et al., 2004). This suggests involvement with U1 snRNA, which recognises this sequence. While U1 snRNA inhibition does cause some stabilisation of PROMPTs, it does not generally result in their longer extension and so other *cis*-acting sequences and processes may additionally contribute (Oh et al., 2017). Our evidence that ZC3H4 binds RNA in cells suggests that it may directly interact with some of its target transcripts and it will be important to delineate any sequence determinants.

Beyond transcriptional regulation, ZC3H4 occupancy of SEs is interesting. Other notable SE-associated factors (e.g. BRD4 and MED1) are much more generally implicated in Pol II transcription than ZC3H4 (Sabari et al., 2018). Moreover, they are transcriptional activators whereas ZC3H4 appears to suppress transcription (or, at least, its RNA output). Many SE-bound factors are found to have phase separation properties explaining why large clusters of factors accumulate at these regions (Cho et al., 2018). While we do not know whether ZC3H4 can phase separate, it contains large regions of intrinsic disorder, which can promote this property (Figure 1 – figure supplement 2A). In general, ZC3H4 may offer a new way to study enhancer clusters, particularly the importance of restricting transcription across these regions.

In conclusion, we have uncovered ZC3H4 as a factor with a function in restricting transcription. Its most notable effects are at non-coding loci where transcriptional termination mechanisms are less understood than at protein-coding genes. Further dissection of ZC3H4 and its targeting should reveal additional important insights into how this unstable portion of the transcriptome is controlled. The non-overlapping effects of

Integrator and ZC3H4 at protein-coding genes indicate the possibility that multiple factors may control gene output via premature transcriptional termination.

ACKNOWLEDGEMENTS

We are grateful to the other members of the lab for critical comment. This work was supported by a Wellcome Trust Investigator Award (WT107791/Z/15/Z) and a Lister Institute Research Fellowship held by S.W. We are grateful to The University of Exeter Sequencing Service where all sequencing was performed who are supported by a Medical Research Council Clinical Infrastructure Award (MR/ M008924/1), the Wellcome Trust Institutional Strategic Support Fund (WT097835MF), a Wellcome Trust Multi User Equipment Award (WT101650MA), and a Biotechnology and Biological Sciences Research Council Longer and Larger (LoLa) Award (BB/ K003240/1).

MATERIALS AND METHODS

Key Resources Table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Cell line (Human)	HCT116-CPSF30-mAID	In-house	This paper	
Cell line (Human)	HCT116-CPSF30-mAID:RPB1-mTurbo	In-house	This paper	
Cell line (Human)	HCT116-ZC3H4-HA-DHFR	In-house	This paper	
Cell line (Human)	HCT116-DIS3-AID	In-house	PMID: 30840897	

Cell line (Human)	HCT116-PNUTS-AID	In-house	This paper	
Cell line (Human)	HCT116-INTS11-SMASh	In-house	PMID: 33113359	
Recombinant DNA reagent	3xHA-mTurbo-NLS_pCDNA3	Addgene	RRID #: Addgene_107172	
Recombinant DNA reagent	px300	Addgene	RRID #: Addgene_42230	
Recombinant DNA reagent	ZC3H4-pcDNA3.1(+)-C-eGFP	Genscript	Custom synthesis	ENTS00000253048
Recombinant DNA reagent	pSL-MS2-6x	Addgene	RRID #: Addgene_27118	
Recombinant DNA reagent	pcDNA3.1(+)IRES GFP	Addgene	RRID #: Addgene_51406	
Antibody	CPSF30	Bethyl	RRID #: AB_2780000 Cat #: A301-585A-T	(1:1000)
Antibody	RNA Pol II	Abcam	RRID #: AB_306327 Cat #: ab817	Now discontinued at abcam (1:1000 for western blot. 4-5ug used for ChIP qPCR and –seq, respectively)

Antibody	PNUTS	Bethyl	RRID #: AB_2779219 Cat #: A300-439A-T	(1:1000)
Antibody	WDR82	Cell Signalling	RRID #: AB_2800319 Clone: D2I3B Cat #: 99715	(1:1000)
Antibody	EXOSC10	Santa Cruz	RRID #: AB_10990273 Cat #: sc-374595	(1:2000)
Antibody	ZC3H4	Atlas Antibodies	RRID #: AB_10795495 Cat #: HPA040934	(1:1000)
Antibody	HA tag	Roche	RRID #: AB_390918 Clone: 3f10 Cat #: 11867423001	(1:2000)
Antibody	GFP	Chromotek	Clone: PABG1 Cat #: PABG1-100 RRID #: AB_2749857	(1:2000)
Antibody	TCF4/TCF7L2	Cell Signalling	RRID #: AB_2199816 Clone: C48H11 Cat #: 2569	(1:1000)
Chemical compound drug	TMP	Sigma	Cat #: T7883	
Chemical compound drug	IAA	Sigma	Cat #: 12886	
Commercial assay, kit	Lipofectamine RNAiMax	Life Technologies	Cat #: 13778075	

Commercial assay, kit	JetPRIME	PolyPlus	Cat #: 114-01	
Commercial assay, kit	Streptavidin Sepharose High Performance slurry	GE Healthcare	Cat #: GE28-9857-38	
Commercial assay, kit	GFP TRAP magnetic agarose	Chromotek	RRID #: AB_2827592 Cat #: gtd-100	
Commercial assay, kit	Dynabeads α-Mouse	Life Technologies	RRID #: AB_2783640 Cat #: 11201D	
Commercial assay, kit	Dynabeads α-Rabbit	Life Technologies	RRID #: AB_2783009 Cat #: 11203D	
Commercial assay, kit	SimpleChIP® Plus Enzymatic Chromatin kit	Cell Signalling	Cat #: 9005	
Commercial assay, kit	TruSeq Stranded Total RNA Library Prep Kit	Illumina	Cat #: 20020596	
Commercial assay, kit	NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®	NEB	Cat #: E7645S	
Commercial assay, kit	Ribo-Zero Gold rRNA removal kit	Illumina	Cat #: 20040526	

Commercial assay, kit	Ampure XP beads	Beckman Coulter	Cat #: A63880	
Commercial assay, kit	RNAClean XP Beads	Beckman Coulter	Cat #: A63987	
Software, algorithm	BamTools	(Barnett et al., 2011)	RRID #: SCR_015987	v2.4.0
Software, algorithm	BEDtools	(Quinlan and Hall, 2010)	RRID #: SCR_006646	v2.26.1
Software, algorithm	Bioconductor	https://bioconductor.org/	RRID #: SCR_006442	v3.11
Software, algorithm	DeepTools	(Ramirez et al., 2014)	RRID #: SCR_016366	v3.3.0
Software, algorithm	DESeq2	(Love et al., 2014)	RRID #: SCR_015687	v1.28.1
Software, algorithm	featureCounts	(Liao et al., 2013, 2014)	RRID #: SCR_012919	v2.0.0
Software, algorithm	FIMO	(Grant et al., 2011)	RRID #: SCR_001783	v5.3.3
Software, algorithm	genomicRanges	http://bioconductor.org/packages/release/bioc/html/GenomicRanges	RRID #: SCR_000025	v1.40.0
Software, algorithm	ggplot2	https://cran.r-project.org/web/packages/ggplot2	RRID #: SCR_014601	v3.3.3

Software, algorithm	Hisat2	(Kim et al., 2015)	RRID #: SCR_015530	v2.1.0
Software, algorithm	IGV	(Robinson et al., 2011)	RRID #: SCR_011793	v2.8.2
Software, algorithm	MACS2	(Zhang et al., 2008)	RRID #: SCR_013291	v2.2.6
Software, algorithm	pheatmap	https://cran.r-project.org/web/packages/pheatmap	RRID #: SCR_016418	v1.0.12
Software, algorithm	R	https://cran.r-project.org/	NA	v4.0.4
Software, algorithm	Rstudio	https://rstudio.com/	RRID #: SCR_000432	v1.3.1093
Software, algorithm	rtracklayer	https://bioconductor.org/packages/release/bioc/html/rtracklayer	NA	v1.48.0
Software, algorithm	SAMTools	(Li et al., 2009)	RRID #: SCR_002105	v.1.11
Software, algorithm	Trim_galore!	https://github.com/FelixKrueger/TrimGalore/	RRID #: SCR_011847	v.0.6.5dev

Cell culture

HCT116 parental cells and engineered cell lines were tested negative for mycoplasma and cultured in Dulbecco modified eagle medium, supplemented with 10% foetal calf serum and penicillin streptomycin (Gibco). For RNAi, 6 or 24-well dishes were transfected with siRNA using Lipofectamine RNAiMax (Life Technologies) following the manufacturers' guidelines. The transfection was repeated 24 hours later and, 48 hours after

the first transfection, RNA was isolated. For MS2 assays, cells were seeded in 24-well dishes overnight, then transfected with 50ng MS2hp-IRES-GFP and 100ng of MS2-GFP, ZC3H4-MS2 or ZC3H4-GFP using JetPRIME® (PolyPlus) for 24 hours. To deplete DIS3-AID or PNUTS-AID, auxin was used at a final concentration of 500uM. To deplete ZC3H4-DHFR, cells were washed twice in PBS and grown in media with or without TMP (30uM).

Cell line generation and cloning

CPSF30-mAID and *CPSF30-mAID*:*RPB1-mTurbo* cells were generated using CRISPR/Cas9-mediated homology-directed repair (HDR). CPSF30 and RPB1 homology arms and gRNA sequences are detailed in Supplementary File 7. The mTurbo insert derives from 3xHA-mTurbo-NLS_pCDNA3 (#107172, Addgene). For ZC3H4 degron cells, 3xHA-DHFR was amplified from existing CPSF73-HA-DHFR constructs (published in (Eaton et al., 2018)) using non-homologous end-joining (NHEJ) as described in (Manna et al., 2019). PNUTS-AID cells were constructed using the protocol described in (Davidson et al., 2019). In general, 6cm dishes of cells were transfected with 1ug of guide RNA expressing px300 plasmid (#42230, Addgene) and 1ug of each HDR template/NHEJ PCR product. Three days later cells were seeded, as appropriate, into hygromycin (30µg/ml, final) neomycin (800µg/ml, final) or puromycin (1µg/ml, final). ZC3H4 cDNA was purchased from Genscript in a pcDNA3.1(+)-C-eGFP vector. The MS2hp-IRES-GFP reporter was made by inserting a BamH1/EcoRV restriction fragment from pSL-MS2-6x (#27118, Addgene) into pcDNA3.1(+)IRES GFP (#51406, Addgene) also digested with BamH1/EcoRV.

Turbo sample preparation

10 cm dishes at ~80% confluency were labelled with 500 µM biotin for 10 mins and the labelling reaction quenched immediately by washing cells in ice cold PBS. Cells were lysed in RIPA buffer (150 mM NaCl, 1% NP40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris-HCl at pH 8, 5 mM EDTA at pH 8) containing protease inhibitors (cOmplete mini EDTA-free tablets, Roche) for 30 mins on ice, then clarified *via* centrifugation. 350 uL of washed Streptavidin Sepharose High Performance slurry (GE Healthcare) was incubated with biotinylated or control lysates with inversion at room temperature for 1 hour. Samples were then washed twice with RIPA buffer, twice with Urea buffer (2 M urea, 50 mM Tris HCl pH 8), twice with 100 mM sodium carbonate and once with (20 mM Tris HCl pH 8, 2 mM $CaCl_2$). Residual final wash buffer was used to re-suspend the beads, which were then flash frozen in liquid nitrogen and sent for Tandem Mass Spectrometry at The University of Bristol Proteomics Facility.

Identifying mass spectrometry candidates

First, contaminant proteins (e.g. keratin) or those that are known to be preferentially biotinylated in ligase assays (e.g. AHNAK) were excluded. Samples with an average abundance ratio of $\leq$ 0.70 were classified as having a decreased interaction with RNA polymerase II following CPSF30 depletion. Finally, proteins with $\leq$ 5 peptides were discarded. Remaining candidates were plotted in Figure 1F.

qRT-PCR

1 µg of total RNA (DNase treated) was reversed transcribed using random hexamers according to manufacturer's instructions (Protoscript II, NEB); cDNA diluted to 50uL. qPCR was performed using LUNA SYBR (NEB) on a Rotorgene (Qiagen). Fold changes were calculated using Qiagen's software based on delta CT values. Graphs were plotted using Prism (GraphPad). Numbers underpinning qPCR-derived bar graphs are provided in source data file 1.

Antibodies

CPSF30 (A301-585A-T, Bethyl), RNA Pol II (ab817, Abcam), PNUTS (A300-439A-T, Bethyl), WDR82 (D2I3B, Cell Signalling), EXOSC10 (sc-374595, Santa Cruz), ZC3H4 (HPA040934, Atlas Antibodies), HA tag (clone 3f10, 11867423001, Roche), GFP (PABG1, Chromotek), TCF4/TCF7L2 (C48H11, Cell Signalling). Uncropped western blots are provided in source data file 2.

GFP Trap

10 cm dishes were transfected (5ug plasmid, 24 hrs), washed with PBS and lysed for 30 mins on ice in 1 mL lysis buffer (150 mM NaCl, 2.5 mM $MgCl_2$, 20 mM Tris HCl pH 7.5, 1 % Triton X-100, 250 units Benzonase [Sigma]). Samples were then clarified through centrifugation (12000xg, 10 mins), split in two and incubated with 25ul of GFP TRAP magnetic agarose (Chromotek) for 1hr with rotation at 4°C. Beads were washed 5x with lysis buffer and samples eluted by boiling in 2xSDS buffer before analysis by western blotting.

Nuclear RNA-Seq

Nuclei were extracted from 1x 30mm dish of cells per condition using hypotonic lysis buffer (10 mM Tris pH5.5, 10 mM NaCl, 2.5 mM $MgCl_2$, 0.5% NP40) with a 10% sucrose cushion and RNA was isolated using Tri-reagent. Following DNase treatment, RNA was Phenol Chloroform extracted and ethanol precipitated. After assaying quality control using a Tapestation (Agilent), 1 µg RNA was rRNA-depleted using Ribo-Zero Gold rRNA removal kit (Illumina) then cleaned and purified using RNAClean XP Beads (Beckman Coulter). Libraries

were prepared using TruSeq Stranded Total RNA Library Prep Kit (Illumina) and purified using Ampure XP beads (Beckman Coulter). A final Tapestation D100 screen was used to determine cDNA fragment size and concentration before pooling and sequencing using Hiseq2500 (Illumina).

ChIP-qPCR

Cells were cross-linked for 10 mins at room temperature (1% Formaldehyde) and quenched for 5 mins (125mM Glycine). Cells were washed in PBS, pelleted (500xg) and resuspended in 400ul RIPA buffer (150 mM NaCl, 1% NP40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris-HCl at pH 8, 5 mM EDTA at pH 8). Sonication was then performed in a Bioruptor (30 seconds on/30 seconds off x10 on high setting) and debris pelleted (13000rpm x 10mins). Supernatants were then incubated for 2hr at 4^{o}C with 40ul of sheep anti-mouse dynabeads to which 4ug of anti-Pol II (or, as a control, nothing) was pre-bound. Beads were washed 6x with RIPA buffer and then bound chromatin was eluted by 30 min incubation at room temperature with rotation (500ul 0.1 M NaHCO3 + 1% SDS). Cross-links were reversed overnight at 65^{o}C with the addition of 20ul 5M NaCl. Following phenol chloroform extraction and ethanol precipitation, chromatin was resuspended in 100ul water of which 1ul was used per qPCR reaction.

ChIP-Seq

ChIP libraries were prepared using SimpleChIP® Plus Enzymatic Chromatin kit (9005, Cell Signalling) according to manufacturer's instructions. 5µg of RNA Pol II (abcam, 8WG16) or ZC3H4 (HPA040934, Atlas Antibodies) were used for immunoprecipitation, Dynabeads α-Mouse / α-Rabbit (Life Technologies) were used for capture.

Chromatin RNA isolation

HCT116 cells were scraped into PBS, pelleted, incubated in hypotonic lysis buffer (HLB; 10 mM Tris.HCl at pH 7.5, 10 mM NaCl, 2.5 mM $MgCl_2$, 0.5% NP40), underlayered with 10% sucrose (*w/v* in HLB) on ice for 5 mins, then spun at 500 xg to isolate nuclei. Supernatant and sucrose was removed and nuclei re-suspended in 100 µL of NUN1 (20 mM Tris-HCl at pH 7.9, 75 mM NaCl, 0.5 mM EDTA, 50% glycerol, 0.85 mM DTT), before being incubated with 1 mL NUN2 (20 mM HEPES at pH 7.6, 1 mM DTT, 7.5 mM $MgCl_2$, 0.2 mM EDTA. 0.3 M NaCl, 1 M urea, 1% NP40) on ice for 15 mins. Samples were spun at 13, 000 xg to pellet chromatin, this was dissolved in Trizol and RNA extracted.

Colony formation assays

ZC3H4-DHFR cells were seeded into 100mm dishes and maintained in the presence or absence of TMP for 10 days, with media replaced every 3 days. Colonies were fixed in ice cold methanol for 10 mins and then stained with 0.5% crystal violet (in 25% methanol) for 10 minutes.

XRNAX

We essentially followed the protocol of (Trendel et al., 2019). HCT116 cells were grown overnight in the presence or absence of doxycycline in 10 cm dishes. 24 hr later, dishes were washed with PBS, UV cross linked (Stratalinker 1800 150 mJ/cm2), or not, then re-suspended in 4.5 mL Trizol (Sigma). 300 uL of chloroform was added, samples agitated on a ThermoMixer (Eppendorf) for 5 mins, span at 12000xg for 15 minutes, then the interphase carefully aspirated into fresh tubes. The interphase was washed thrice with Tris-SDS (10 mM Tris-HCL pH 7.5, 1 mM EDTA, 0.1 % SDS), before being dissolved in 1 mL Tris-SDS. 1 uL glycogen, 60 uL of 5M NaCl and 1 mL isopropanol were added and samples precipitated at -20'C for 10 minutes, then pelleted at 18, 000xg for 15 mins. Precipitated protein was washed with 70 % ethanol, air dried, re-suspended in 180 uL water and pellets dissolved on ice. DNA was removed via TurboDNase (ThermoFisher) treatment, before samples were re-pelleted, re-dissolved in RNase buffer (150 mM NaCl, 20 mM Tris-HCL pH 7.5, 2.5 mM MgCl2) and RNA digested with RNase H (NEB) and 1 uL of RNAse T1 (Roche). 4 x SDS loading buffer was added before gel electrophoresis and western blotting.

Computational analysis

All sequencing data were uploaded to the Galaxy web platform and processed as detailed below; usegalaxy.org and usegalaxy.eu servers were used.

Datasets (GEO accessions)

Data newly generated in this paper (GSE163015); Pol II HEPG2 ChIP-seq (GSE32883); ZC3H4 HEPG2 ChIP-seq (GSE104247); DIS3-AID HCT116 RNA-seq (GSE120574); INTS1 RNAi Chromatin-associated RNA-seq (GSE150238). 4sU labelled RNA differential expression in HeLa cells depleted of INTS11 or ZC3H4 (GSE133109, GSE151919).

RNA-Seq alignment

FASTA files were trimmed using Trim Galore! and mapped to GRCh38 using HISAT2 using default parameters (Kim et al., 2015). Reads with a MAPQ score of $\leq$ 20 were removed from alignment files using SAMtools (Li et al., 2009). Finally, BigWig files were generated using DeepTools and visualised using IGV (Ramirez et al., 2014).

ChIP alignment and visualisation

All samples were mapped against GRCh38 using BWA, default settings. Reads with a MAPQ score of ≤ 20 were removed along with PCR duplicates from alignment files using SAMtools. Processed BAM files were converted to BigWig files using DeepTools: all samples were normalised to RPKM with a bin size of 1. Aligned files were visualised using IGV.

ChIP peak calling

For ZC3H4 and INPUT, broad peaks were called separately using MACS2 with a changed “lower mfold” (2) and default settings. For each experiment, bedtools was used to establish common peaks from individual reps (Intersect Intervals), creating a bed file of high confidence peaks. For ZC3H4, peaks called in the INPUT sample were subtracted *via* bedtools. All bed files were annotated and plotted in R using ChipSeeker (Yu et al., 2015).

Gene heatmaps

For ChIP heatmaps, computematrix (DeepTools) was used to generate score files from ChIP bigwig files using an hg38 bed file; parameters used for each heat map are detailed in figure legends. Plots were redrawn in R. Transcription read-through analysis was calculated for each condition by comparing the first 1 kb downstream of the TES to a 500 bp region directly preceding the TES (PAS). A log2 ratio (depletion/control) was then applied to identify increased read-through.

Super-enhancer metaplots

A bed file with the coordinates of super-enhancer locations from dbSUPER in HCT116 cells was used as a basis (Khan and Zhang, 2016). All regions that had clusters of MED1, Pol II and H3K27ac ChIP signal were retained as *bone fide* regions of interest, those without were discarded. A log2 ratio of experiment vs input was prepared using BamCompare of DeepTools - for RNAseq metaplots, BAM files were split by strand. A score file for the regions in the amended SE bed file was generated *via* the computematrix function of DeepTools using the log2 BamCompare output file. Results were plotted in R-studio using ggplot2.

Gene plots and metaplots

Split strand metagene plots were generated using RPKM normalised sense and antisense (scaled to -1) bigwig coverage files separately with further graphical processing performed in R. For identifying ZC3H4 PROMPT regions ncRNA genes were filtered from hg38 refgene gtf file to give protein-coding genes that were used with feature counts on

siCont RNAseq (Liao et al., 2014), to gain read count and gene length. Transcripts per million (TPM) were calculated for each gene and genes with an expression of < 5 were filtered out to give a list of expressed genes. Next, divergent promoters, or genes with neighbours within 5kb of their promoter, were excluded to minimise background. Finally, this gene list was converted to a bed file, then computematrix (DeepTools) used to generate a score file from log2 siCont Vs condition bigwigs; results were plotted in R.

Differential gene expression

FeatureCounts was used to count mapped reads over exons and differential expression was performed using DESeq2 (Liao et al., 2014; Love et al., 2014).

PROMPT poly-A site detection

For PROMPT analysis, we used a catalogue of 961 PROMPTs generated by *de novo* assembly following acute DIS3 depletion (Davidson et al., 2019). Due to the variable length of each PROMPT, we searched for the two consensus poly-A site motifs (AWTAAA) across the full transcript sequence using FIMO (online). We then calculated the total occurrence of poly-A sites across each PROMPT transcript per kb and separated them into two groups; those that show upregulation (log2FC ≥1) in the absence of ZC3H4 and those with no change of downregulated expression. Plots were drawn in R.

ZC3H4 homologue identification

To identify ZC3H4 homolog protein sequences, sequences from UniRef100 (UniProt Consortium, 2014) were searched using a profile HMM search: 'hmmsearch', part of HMMer V3.2.1 (Eddy, 2011), with PANTHER (Mi et al., 2019) hidden Markov model PTHR13119, corresponding to zinc finger CCCH-domain containing proteins. Profile HMM search hits were filtered using a 1e-100 e-value threshold; this search identified 1513 UniRef100 sequences with PTHR13119 domains (representing a total of 1646 UniProtKB sequences). PTHR13119 domains from human and mouse were aligned using TCoffee Expresso mode (Armougom et al., 2006), and multiple sequence alignment figure (Figure 1 – figure supplement 3B) was rendered with ESPscript (Robert and Gouet, 2014).

Phylogenetic tree reconstruction

Identified PTHR13119 domains were aligned using MAFFT v7.4 (Katoh and Standley, 2013); sites composed of more than 75% of gaps were removed from the multiple sequence alignment with trimAl (Capella-Gutierrez et al., 2009). The PTHR13119 domain phylogeny was reconstructed under maximum likelihood with IQ-TREE v1.6.9 (Nguyen et al.,

2015). The best-fitting substitution matrix was determined by ModelFinder (Kalyaanamoorthy et al., 2017), as implemented in IQ-TREE. Branch support values were based on 1000 ultrafast bootstraps (Minh et al., 2013). Phylogenetic Tree figure was rendered with iToL (Letunic and Bork, 2019). Multiple sequence alignment and phylogenetic tree files are deposited on Zenodo (https://doi.org/10.5281/zenodo.4637127).

Primers, siRNAs and other nucleic acid sequences

Sequences are provided in Supplementary File 7.

REFERENCES

Addepalli, B., and Hunt, A.G. (2007). A novel endonuclease activity associated with the Arabidopsis ortholog of the 30-kDa subunit of cleavage and polyadenylation specificity factor. Nucleic acids research *35*, 4453-4463.

Andersson, R., Refsing Andersen, P., Valen, E., Core, L.J., Bornholdt, J., Boyd, M., Heick Jensen, T., and Sandelin, A. (2014). Nuclear stability and transcriptional directionality separate functionally distinct RNA species. Nat Commun *5*, 5336.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C. (2006). Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic acids research *34*, W604-608.

Austenaa, L.M., Barozzi, I., Simonatto, M., Masella, S., Della Chiara, G., Ghisletti, S., Curina, A., de Wit, E., Bouwman, B.A., de Pretis, S.*, et al.* (2015). Transcription of Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early Termination. Molecular cell *60*, 460-474.

Austenaa, L.M.I., Piccolo, V., Russo, M., Prosperini, E., Polletti, S., Polizzese, D., Ghisletti, S., Barozzi, I., Diaferia, G.R., and Natoli, G. (2021). A first exon termination checkpoint preferentially suppresses extragenic transcription. Nat Struct Mol Biol.

Bai, C., and Tolias, P.P. (1996). Cleavage of RNA hairpins mediated by a developmentally regulated CCCH zinc finger protein. Molecular and cellular biology *16*, 6661-6667.

Baillat, D., Hakimi, M.A., Naar, A.M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. Cell *123*, 265-276.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stromberg, M.P., and Marth, G.T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691-1692.

Beckedorff, F., Blumenthal, E., daSilva, L.F., Aoi, Y., Cingaram, P.R., Yue, J., Zhang, A., Dokaneheifard, S., Valencia, M.G., Gaidosh, G.*, et al.* (2020). The Human Integrator

730 Complex Facilitates Transcriptional Elongation by Endonucleolytic Cleavage of Nascent
731 Transcripts. Cell reports *32*, 107917.

732 Branon, T.C., Bosch, J.A., Sanchez, A.D., Udeshi, N.D., Svinkina, T., Carr, S.A., Feldman,
733 J.L., Perrimon, N., and Ting, A.Y. (2018). Efficient proximity labeling in living cells and
734 organisms with TurboID. Nat Biotechnol *36*, 880-887.

735 Brewer-Jensen, P., Wilson, C.B., Abernethy, J., Mollison, L., Card, S., and Searles, L.L.
736 (2016). Suppressor of sable [Su(s)] and Wdr82 down-regulate RNA from heat-shock-
737 inducible repetitive elements by a mechanism that involves transcription termination. Rna *22*,
738 139-154.

739 Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for
740 automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*,
741 1972-1973.

742 Chan, S.L., Huppertz, I., Yao, C., Weng, L., Moresco, J.J., Yates, J.R., 3rd, Ule, J., Manley,
743 J.L., and Shi, Y. (2014). CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA
744 3' processing. Genes & development *28*, 2370-2380.

745 Chiu, A.C., Suzuki, H.I., Wu, X., Mahat, D.B., Kriz, A.J., and Sharp, P.A. (2018).
746 Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature
747 Polyadenylation Suppressed by U1 snRNP. Molecular cell *69*, 648-663 e647.

748 Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, II (2018). Mediator
749 and RNA polymerase II clusters associate in transcription-dependent condensates. Science
750 *361*, 412-415.

751 Chung, H.K., Jacobs, C.L., Huo, Y., Yang, J., Krumm, S.A., Plemper, R.K., Tsien, R.Y., and
752 Lin, M.Z. (2015). Tunable and reversible drug control of protein production via a self-excising
753 degron. Nature chemical biology *11*, 713-720.

754 Clerici, M., Faini, M., Muckenfuss, L.M., Aebersold, R., and Jinek, M. (2018). Structural basis
755 of AAUAAA polyadenylation signal recognition by the human CPSF complex. Nat Struct Mol
756 Biol *25*, 135-138.

757 Cortazar, M.A., Sheridan, R.M., Erickson, B., Fong, N., Glover-Cutter, K., Brannan, K., and
758 Bentley, D.L. (2019). Control of RNA Pol II Speed by PNUTS-PP1 and Spt5
759 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism.
760 Molecular cell.

761 Davidson, L., Francis, L., Cordiner, R.A., Eaton, J.D., Estell, C., Macias, S., Caceres, J.F.,
762 and West, S. (2019). Rapid Depletion of DIS3, EXOSC10, or XRN2 Reveals the Immediate
763 Impact of Exoribonucleolysis on Nuclear RNA Metabolism and Transcriptional Control. Cell
764 reports *26*, 2779-2791 e2775.

765 Davidson, L., Francis, L., Eaton, J.D., and West, S. (2020). Integrator-Dependent and
766 Allosteric/Intrinsic Mechanisms Ensure Efficient Termination of snRNA Transcription. Cell
767 reports *33*, 108319.

768 Dubbury, S.J., Boutz, P.L., and Sharp, P.A. (2018). CDK12 regulates DNA repair genes by
769 suppressing intronic polyadenylation. Nature *564*, 141-145.

770 Dye, M.J., Gromak, N., and Proudfoot, N.J. (2006). Exon tethering in transcription by RNA
771 polymerase II. Molecular cell *21*, 849-859.

772 Eaton, J.D., Davidson, L., Bauer, D.L.V., Natsume, T., Kanemaki, M.T., and West, S. (2018).
773 Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73
774 activity. Genes & development *32*, 127-139.

775 Eaton, J.D., Francis, L., Davidson, L., and West, S. (2020). A unified allosteric/torpedo
776 mechanism for transcriptional termination on human protein-coding genes. Genes &
777 development *34*, 132-145.

778 Eaton, J.D., and West, S. (2020). Termination of Transcription by RNA Polymerase II:
779 BOOM! Trends Genet *36*, 664-675.

780 Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, e1002195.

781 Elrod, N.D., Henriques, T., Huang, K.L., Tatomer, D.C., Wilusz, J.E., Wagner, E.J., and
782 Adelman, K. (2019). The Integrator Complex Attenuates Promoter-Proximal Transcription at
783 Protein-Coding Genes. Molecular cell *76*, 738-752 e737.

784 Fong, N., Brannan, K., Erickson, B., Kim, H., Cortazar, M.A., Sheridan, R.M., Nguyen, T.,
785 Karp, S., and Bentley, D.L. (2015). Effects of Transcription Elongation Rate and Xrn2
786 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic
787 Competition. Molecular cell *60*, 256-267.

788 Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given
789 motif. Bioinformatics *27*, 1017-1018.

790 Gregersen, L.H., Mitter, R., Ugalde, A.P., Nojima, T., Proudfoot, N.J., Agami, R., Stewart, A.,
791 and Svejstrup, J.Q. (2019). SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. Cell *177*,
792 1797-1813 e1718.

793 Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and
794 Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*,
795 934-947.

796 Iasillo, C., Schmid, M., Yahia, Y., Maqbool, M.A., Descostes, N., Karadoulama, E., Bertrand,
797 E., Andrau, J.C., and Jensen, T.H. (2017). ARS2 is a general suppressor of pervasive
798 transcription. Nucleic acids research *45*, 10229-10241.

799 Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P.,
800 Roth, A., Simonovic, M.*, et al.* (2009). STRING 8--a global view on proteins and their
801 functional interactions in 630 organisms. Nucleic acids research *37*, D412-416.

Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664-668.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermiin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods *14*, 587-589.

Kamieniarz-Gdula, K., Gdula, M.R., Panser, K., Nojima, T., Monks, J., Wisniewski, J.R., Riepsaame, J., Brockdorff, N., Pauli, A., and Proudfoot, N.J. (2019). Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. Molecular cell *74*, 158-172 e159.

Kamieniarz-Gdula, K., and Proudfoot, N.J. (2019). Transcriptional Control by Premature Termination: A Forgotten Mechanism. Trends Genet *35*, 553-564.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol *30*, 772-780.

Khan, A., and Zhang, X. (2016). dbSUPER: a database of super-enhancers in mouse and human genome. Nucleic acids research *44*, D164-171.

Kieft, R., Zhang, Y., Marand, A.P., Moran, J.D., Bridger, R., Wells, L., Schmitz, R.J., and Sabatini, R. (2020). Identification of a novel base J binding protein complex involved in RNA polymerase II transcription termination in trypanosomes. PLoS Genet *16*, e1008390.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat Methods *12*, 357-360.

Kopp, F., and Mendell, J.T. (2018). Functional Classification and Experimental Dissection of Long Noncoding RNAs. Cell *172*, 393-407.

Kuan, Y.S., Brewer-Jensen, P., and Searles, L.L. (2004). Suppressor of sable, a putative RNA-processing protein, functions at the level of transcription. Molecular and cellular biology *24*, 3734-3746.

Kustatscher, G., Grabowski, P., Schrader, T.A., Passmore, J.B., Schrader, M., and Rappsilber, J. (2019). Co-regulation map of the human proteome enables identification of protein functions. Nat Biotechnol *37*, 1361-1371.

Lai, F., Gardini, A., Zhang, A., and Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. Nature *525*, 399-403.

Lee, J.H., and Skalnik, D.G. (2008). Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed human genes. Molecular and cellular biology *28*, 609-618.

Lee, J.H., You, J., Dobrota, E., and Skalnik, D.G. (2010). Identification and characterization of a novel human PP1 phosphatase complex. J Biol Chem *285*, 24466-24476.

838 Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new
839 developments. Nucleic acids research *47*, W256-W259.

840 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis,
841 G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence
842 Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

843 Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable
844 read mapping by seed-and-vote. Nucleic acids research *41*, e108.

845 Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose
846 program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930.

847 Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and
848 dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

849 Lykke-Andersen, S., Zumer, K., Molska, E.S., Rouviere, J.O., Wu, G., Demel, C., Schwalb,
850 B., Schmid, M., Cramer, P., and Jensen, T.H. (2020). Integrator is a genome-wide attenuator
851 of non-productive transcription. Molecular cell.

852 Manna, P.T., Davis, L.J., and Robinson, M.S. (2019). Fast and cloning-free CRISPR/Cas9-
853 mediated genomic editing in mammalian cells. Traffic *20*, 974-982.

854 Mendoza-Figueroa, M.S., Tatomer, D.C., and Wilusz, J.E. (2020). The Integrator Complex in
855 Transcription and Development. Trends Biochem Sci *45*, 923-934.

856 Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version
857 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis
858 tools. Nucleic acids research *47*, D419-D426.

859 Minh, B.Q., Nguyen, M.A., and von Haeseler, A. (2013). Ultrafast approximation for
860 phylogenetic bootstrap. Mol Biol Evol *30*, 1188-1195.

861 Natsume, T., Kiyomitsu, T., Saga, Y., and Kanemaki, M.T. (2016). Rapid Protein Depletion in
862 Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. Cell reports
863 *15*, 210-218.

864 Nedea, E., He, X., Kim, M., Pootoolal, J., Zhong, G., Canadien, V., Hughes, T., Buratowski,
865 S., Moore, C.L., and Greenblatt, J. (2003). Organization and function of APT, a subcomplex
866 of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and
867 small nucleolar RNA 3'-ends. J Biol Chem *278*, 33000-33010.

868 Nguyen, L.T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and
869 effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol
870 *32*, 268-274.

871 Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M.,
872 and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent
873 Transcription Coupled to RNA Processing. Cell *161*, 526-540.

874 Nojima, T., Tellier, M., Foxwell, J., Ribeiro de Almeida, C., Tan-Wong, S.M., Dhir, S.,
875 Dujardin, G., Dhir, A., Murphy, S., and Proudfoot, N.J. (2018). Deregulated Expression of
876 Mammalian lncRNA through Loss of SPT6 Induces R-Loop Formation, Replication Stress,
877 and Cellular Senescence. Molecular cell *72*, 970-984 e977.

878 O'Reilly, D., Kuznetsova, O.V., Laitem, C., Zaborowska, J., Dienstbier, M., and Murphy, S.
879 (2014). Human snRNA genes use polyadenylation factors to promote efficient transcription
880 termination. Nucleic acids research *42*, 264-275.

881 Oh, J.M., Di, C., Venters, C.C., Guo, J., Arai, C., So, B.R., Pinto, A.M., Zhang, Z., Wan, L.,
882 Younis, I.*, et al.* (2017). U1 snRNP telescripting regulates a size-function-stratified human
883 genome. Nat Struct Mol Biol *24*, 993-999.

884 Partridge, E.C., Chhetri, S.B., Prokop, J.W., Ramaker, R.C., Jansen, C.S., Goh, S.T.,
885 Mackiewicz, M., Newberry, K.M., Brandsmeier, L.A., Meadows, S.K.*, et al.* (2020).
886 Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature *583*,
887 720-728.

888 Pott, S., and Lieb, J.D. (2015). What are super-enhancers? Nature genetics *47*, 8-12.

889 Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano,
890 C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription
891 upstream of active human promoters. Science *322*, 1851-1854.

892 Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes &
893 development *25*, 1770-1782.

894 Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing
895 genomic features. Bioinformatics *26*, 841-842.

896 Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A., and Manke, T. (2014). deepTools: a
897 flexible platform for exploring deep-sequencing data. Nucleic acids research *42*, W187-191.

898 Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the new
899 ENDscript server. Nucleic acids research *42*, W320-324.

900 Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and
901 Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24-26.

902 Sabari, B.R., Dall'Agnese, A., Boija, A., Klein, I.A., Coffey, E.L., Shrinivas, K., Abraham, B.J.,
903 Hannett, N.M., Zamudio, A.V., Manteiga, J.C.*, et al.* (2018). Coactivator condensation at
904 super-enhancers links phase separation and gene control. Science *361*.

905 Schlackow, M., Nojima, T., Gomes, T., Dhir, A., Carmo-Fonseca, M., and Proudfoot, N.J.
906 (2017). Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs.
907 Molecular cell *65*, 25-38.

908 Sheridan, R.M., and Bentley, D.L. (2016). Selectable one-step PCR-mediated integration of
909 a degron for rapid depletion of endogenous human proteins. Biotechniques *60*, 69-74.

Sun, Y., Zhang, Y., Hamilton, K., Manley, J.L., Shi, Y., Walz, T., and Tong, L. (2018). Molecular basis for the recognition of the human AAUAAA polyadenylation signal. Proceedings of the National Academy of Sciences of the United States of America *115*, E1419-E1428.

Tatomer, D.C., Elrod, N.D., Liang, D., Xiao, M.S., Jiang, J.Z., Jonathan, M., Huang, K.L., Wagner, E.J., Cherry, S., and Wilusz, J.E. (2019). The Integrator complex cleaves nascent mRNAs to attenuate transcription. Genes & development *33*, 1525-1538.

Trendel, J., Schwarzl, T., Horos, R., Prakash, A., Bateman, A., Hentze, M.W., and Krijgsveld, J. (2019). The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. Cell *176*, 391-403 e319.

van Nuland, R., Smits, A.H., Pallaki, P., Jansen, P.W., Vermeulen, M., and Timmers, H.T. (2013). Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. Molecular and cellular biology *33*, 2067-2077.

Wuarin, J., and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Molecular and cellular biology *14*, 7219-7225.

Yu, G., Wang, L.G., and He, Q.Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics *31*, 2382-2383.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W*., et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

FIGURE LEGENDS

Figure 1: Proximity labelling of CPSF30-sensitive Pol II interactions by mTurbo

a) Schematic of the strategy used to tag CPSF30 with the mini auxin-inducible degron (mAID). Guide RNA-expressing Cas9 plasmid and homology-directed repair (HDR) plasmids are shown and the resulting modification to *CPSF30* is represented with each inserted element labelled.

b) Western blot demonstrating CPSF30 depletion. Parental HCT116-TIR1, or *CPSF30-mAID* cells, were treated +/- auxin for 3 hours then blotted. CPSF30 protein is indicated together with a non-specific product, marked by an asterisk, used as a proxy for protein loading.

c) Metagene analysis of 1795 protein-coding genes demonstrating increased downstream transcription, derived from sequencing nuclear RNA, following auxin treatment (3hr) of *CPSF30-mAID* cells. TSS = transcription start site, TES = transcription end site (PAS), read-through signal is normalised against gene body. RPKM is Reads Per Kilobase of transcript,

per Million mapped reads. Positive and negative signals represent sense and antisense reads, respectively.

d) Schematic of our strategy to identify new factors involved in transcription termination. *CPSF30-mAID* cells edited to contain Rpb1-mTurbo (blue circle on Pol II). The addition of biotin induces mTurbo-mediated biotinylation (orange haze) of factors proximal to Pol II. CPSF complex is shown as an example of what might be captured by this experiment.

e) Western blot showing streptavidin HRP probing of extracts from *CPSF30-mAID: RPB1-mTurbo* cells. Prior treatment with auxin (3hr)/biotin (10 mins) is indicated. The high molecular weight species in the + biotin samples corresponds in size to Rpb1-mTurbo (*).

f) Heat map detailing proteins with the largest decrease in Pol II interaction. Data underpinning heat map are from mass spectrometry analysis of streptavidin sequestered peptides (+/- CPSF30) performed in triplicate. Labelling was for 10 minutes.

Figure 2: ZC3H4 depletion stabilises unproductive transcripts.

a) IGV track of the transcription read-through defect at *PTPN11* following CPSF30 or ZC3H4 depletion. Blue and red tracks indicate sense/anti-sense transcripts respectively, grey bar indicates a change in y-axis scale so that comparatively weaker read-through signals can be visualised next to the gene body (left scale for upstream of TES; right for downstream). Y-axis scale is RPKM.

b) Metagene comparison of transcription upstream, across, and downstream of, protein-coding genes in nuclear RNA from *CPSF30-mAID* cells treated or not with auxin and from HCT116 cells transfected with control or ZC3H4 siRNAs. CPSF30 traces are from the same samples presented in Figure 1C. Positive and negative signals represent sense and antisense reads, respectively.

c) IGV track view of transcription at the *MYC* PROMPT in RNA-seq samples obtained from control or ZC3H4 siRNA treated HCT116 cells. We also show a track from HCT116 cells acutely depleted of DIS3-AID (DIS3 +IAA) (Davidson et al., 2019) to highlight the normal extent of this unstable transcript. Y-axis scale is RPKM.

d) Log2 fold change of siZC3H4 vs siControl or DIS3 + vs - auxin for RNA upstream of 6057 non-neighbouring, actively transcribed genes, plotted as heat maps. Line graphs are an XY depiction of heat map data. Log2 fold changes are smaller in siZC3H4 samples versus DIS3 depletion because this is an average of all genes in the heat map, a smaller fraction of which are affected by ZC3H4.

e) IGV plot of a known SE upstream of *MYC* (the location is shown by blue bar under trace). Samples are shown from HCT116 cells treated with control or ZC3H4 siRNAs as well as *DIS3-AID* cells treated with auxin (the latter from (Davidson et al., 2019)) to show the normal extent of unstable eRNAs over this region. Y-axis scale is RPKM.

f) Log2 fold change of RNA signal for siZC3H4 vs siControl or DIS3 + vs - auxin for 111 SEs. The bed file detailing super-enhancer coordinates in HCT116 cells was taken from dbSUPER.org.

Figure 3: Comparison of ZC3H4 and Integrator effects

a) Metagene analysis of chromatin-associated RNA-seq performed on cells treated with control or INTS1-specific siRNA. The plot shows signals upstream, across and downstream of protein-coding genes. Y-axis scale is RPKM. Positive and negative values represent sense and antisense reads, respectively.

b/c) IGV traces of *HAP1* and *NWD1* genes derived from chromatin-associated RNA-seq in control and INTS1 siRNA treated samples and nuclear RNA-seq from control or ZC3H4 siRNA treatment. *NWD1* transcripts are affected by ZC3H4 but not INTS1 whereas the opposite is true for *HAP1* RNAs. Y-axes scales are RPKM.

d) Venn diagram showing the number of mRNAs upregulated ≥ 2-fold, padj ≤0.05 following ZC3H4 depletion versus INTS1 loss and the overlap between the two sets. Genes that showed increased expression due to transcription read-through from an upstream gene were also discarded by assessing coverage over a 1 kb region preceding the TSS, relative to untreated cells. Gene lists are provided in Supplemental File 3.

e) Graphs demonstrating the expression level of mRNA transcripts upregulated (log2FC >1) following ZC3H4 or INTS1 depletion by comparison with transcripts unaffected by loss of either factor. Y-axis shows normalised gene counts (i.e. expression level).

f) Comparison of chromatin-associated RNA-seq in control and INTS1 siRNA treated samples with nuclear RNA-seq derived from control or ZC3H4 siRNA treatment. The *ITPRID2* PROMPT is displayed and y-axes are RPKM (note the different scales between ZC3H4 and INTS1 samples).

g) Comparison of chromatin-associated RNA-seq in control and INTS1 siRNA treated samples with nuclear RNA-seq derived from control or ZC3H4 siRNA treatment. The *MSRB3* SE is displayed and y-axes are RPKM (note the different scales between INTS1 and ZC3H4 samples).

h) Metaplot of RNA-seq profile over super-enhancers following INTS1 depletion (log2 fold depletion/control over 111 super-enhancer as line graphs). The bed file detailing super-enhancer coordinates in HCT116 cells was taken from dbSUPER.org.

Figure 4: Transcriptional dysregulation following acute ZC3H4 loss.

a) Schematic detailing how the DHFR degron works. *E. Coli* dihydrofolate reductase (DHFR) is fused to the C-terminus of ZC3H4, which is stabilised by trimethoprim (TMP). When TMP is removed ZC3H4-DHFR is degraded.

b) Western blot of HCT116 parental and HCT116 *ZC3H4-DHFR* cells +/- TMP. TMP was withdrawn for 4 hours or overnight, EXOCS10 is used as a loading control, αHA recognises a HA peptide before the DHFR tag, while αZC3H4 recognises native protein.

c) Western blot of *ZC3H4-DHFR* cells grown under the following conditions: +TMP, -TMP (4hr), -TMP (4hr) followed by +TMP (4hr). ZC3H4-DHFR is detected using αHA and EXOSC10 is shown as a loading control.

d) qRT-PCR analysis of PROMPT and SE transcripts in *ZC3H4-DHFR* cells grown under the conditions represented in c) (rescue refers to –TMP then +TMP for re-establishing ZC3H4). Graph shows fold change versus +TMP following normalisation to spliced actin. N=3. Error bars are standard error of the mean (SEM). *. ** and *** denote p values of <0.05, 0.01 and 0.001 respectively. ITPRID2 5’ and 3’ primers are at approximately -500bp and -7kb relative to its TSS. HMGA2 5’ and 3’ primers are at approximately -1.8kb and -7.1kb relative to its TSS.

e) qRT-PCR analysis of spliced PJVK, ENO3 and NWD1 mRNAs and RNU1-1 read-through (RT) in *ZC3H4-DHFR* cells grown with or without (4hr) TMP and *INTS11-SMASh* cells grown with or without asunaprevir (ASN; 36 hrs). Graph shows fold change versus control (+TMP for ZC3H4-DHFR samples and –ASN for INTS11-SMASh samples), following normalisation to spliced actin. N=3. Error bars are SEM. * and ** denote p values of <0.05 and 0.01 respectively.

f) Western blot of extracts derived from HCT116 cells transfected with control or WDR82-specific siRNAs. The blot shows WDR82 and, as a loading control, EXOSC10.

g) qRT-PCR of PROMPT and SE transcripts in *ZC3H4-DHFR* cells transfected with control or WDR82 siRNAs before withdrawal, or not, of TMP (14hr). Graph shows fold change by comparison with control siRNA transfected *ZC3H4-DHFR* cells maintained in TMP following

normalisation to spliced actin transcripts. N=3. Error bars are SEM. * and ** denote p<0.05 and 0.01, respectively.

Figure 5: ZC3H4 occupies regions where transcription is affected by its absence.

a) ZC3H4 ChIP profile over protein-coding genes is similar to Pol II. Heat map representation of ZC3H4 and Pol II ChIP-seq occupancy over the gene body +/- 3 kb.

b) ZC3H4 occupies fewer promoters than Pol II. IGV track view of ZC3H4 and Pol II occupancy over *KAZALD1* and *FABP5* genes, Pol II is present at both genes, while ZC3H4 is only present at *KAZALD1*. Scale is counts per million (CPM). Shaded blue box shows peak of Pol II and ZC3H4 at *KAZALD1* and of Pol II over *FABP5*.

c) RNA-seq (HCT116 cells treated with control or ZC3H4 siRNA) and ChIP-seq (Pol II, ZC3H4 and input) profiles at *RPL13.* ZC3H4 occupancy is focused more on the PROMPT transcript region (blue box) than the TSS where, in contrast, the Pol II signal is maximal. RNA-seq scale is RPKM and ChIP-seq is CPM.

d) ZC3H4 ChIP occupancy mirrors Pol II at super-enhancers. IGV track view of ZC3H4 and Pol II occupancy over the SE at the *MSRB3* locus. HCT116 super-enhancer gene track is from dbSUPER and depicted as blue bars.

e) Log2 fold change of ZC3H4 and Pol II vs input at SEs shown as a line graph. Halo denotes 95% confidence level.

f) ChIPseeker analysis of peak distribution of ZC3H4 and Pol II. Occupancy regions are colour-coded and the number of ChIP peaks expressed as a proportion of 100%.

g) Heat map showing Pol II and ZC3H4 ChIP occupancy in HEPG2 cells obtained via the ENCODE consortium. Occupancy +/- 2 kb of the TSS is shown.

Figure 6: Directed recruitment of ZC3H4 recapitulates its effects on endogenous targets

a) Schematic of the MS2 system. A reporter plasmid (MS2hp-IRES-GFP) expressing a GFP transcript with 6 x MS2 hairpins upstream of an IRES and GFP gene. ZC3H4-MS2 or MS2-GFP can be specifically tethered to the MS2 hairpins to assess consequent effects on transcription/RNA output. Positions of primer pairs used in qRT-PCR experiments elsewhere in the figure are indicated by labelled horizontal lines under reporter. POI is protein of interest.

b) qRT-PCR analysis of total RNA isolated from MS2hp-IRES-GFP transfected cells co-transfected with either MS2-GFP, ZC3H4-GFP or ZC3H4-MS2. The level of reporter RNA is plotted ("UP" amplicon) as a percentage of that obtained in the MS2-GFP sample following normalisation to spliced actin. N=3. Error bars are SEM. * denotes $p<0.05$.

c) qRT-PCR analysis of chromatin-associated RNA isolated from MS2hp-IRES-GFP transfected cells co-transfected with either MS2-GFP or ZC3H4-MS2. The level of reporter RNA upstream of the MS2 hairpins (UP) and transcripts yet to be cleaved at the BGH poly(A) site (BGH UC) are plotted as a percentage of that obtained in the MS2-GFP sample following normalisation to spliced actin. N=3. Error bars are SEM. * denotes $p<0.05$.

d) qRT-PCR analysis of total RNA isolated from MS2hp-IRES-GFP transfected *DIS3-AID* cells co-transfected with either MS2-GFP or ZC3H4-MS2 – simultaneously treated or not with auxin (14hr in total). The graph shows the ratio of RNA species recovered upstream (UP) versus downstream (DOWN) of the MS2 hairpins. N=4. Error bars are SEM. * denotes $p<0.05$.

e) Schematic detailing an interplay between ZC3H4 and DIS3 that sees transcription stop and nascent RNA degraded

f) Colony formation assay of *ZC3H4-DHFR* cells grown in the presence or absence of TMP. Cells were grown for 10 days before crystal violet staining.

SUPPLEMENTAL FIGURE LEGENDS

Figure 1 – figure supplement 1

a) IGV track views of the transcription termination defect at *PCBP1*, *PSMC2, LSM8* and *CAV2* genes in the presence (CPSF30-IAA) or absence of (CPSF30 +IAA) in *CPSF30-mAID* cells. Signal is RPKM.

b) Western blot demonstrating bi-allelic modification of RPB1 (Pol II) with mTurbo. The clone employed in Figure 1 is shown against cells parental HCT116 cells unmodified at *RBP1*. The upshift of Pol II signal shows the bi-allelic modification of RPB1. EXOSC10 serves as a loading control.

c) qRT-PCR of total RNA isolated from *CPSF30-mAID: RPB1- mTurbo* cells treated or not with auxin (3hr). An amplicon located ~10kb downstream of the *HMGA2* PAS was used to assay transcriptional read-through presented as a fold change versus minus auxin after normalising to spliced actin mRNA. n=2. Individual data points are shown.

Figure 1 – figure supplement 2

a) Schematic of ZC3H4 and ZC3H6 showing the three CCCH zinc finger domains. A predictor of natural disordered region (PONDR) analysis shows that the only ordered region coincides with these domains. Graph generated via PONDR.com, set to VSL2.

b) STRING analysis of ZC3H4 and ZC3H6 indicate interactors with 3' end processing complex. Image was taken from string.db.org, confidence value was set to medium (0.4). The thickness of lines between nodes is indicative of the confidence in interaction.

c) Proteins that are co-regulated with ZC3H4 according to ProteomeHD using a score cut-off set to 0.98. Table shows GO term analysis of the potentially co-regulated factors.

d) Co-immunoprecipitation of WDR82 using ZC3H4-GFP as bait. Blot shows input (5%) and immunoprecpitated material probed with antibodies to WDR82 or GFP. Cells untransfected with ZC3H4-GFP act as a negative control.

Figure 1 – figure supplement 3

a) Maximum-likelihood phylogenetic tree of zinc finger CCCH-domains (1513 sequences; 795 parsimony informative sites) inferred under the JTT+R8 model. Clades of ZC3H4-like and 845 ZC3H6-like domains are delimited by dashed lines. CCCH-domains identified using the PANTHER hidden Markov model PTHR13119 against the UniProtKB protein database (non redundant version: UniRef100; external node size represents protein cluster size). Branch support values ≥90% (based on 1000 ultrafast bootstraps) are indicated by grey circles. Red stars show SwissProt reviewed protein sequences; external nodes are color-coded according to their taxonomic lineage. Scale bar represents the number of estimated substitutions per site. Virtually all recovered sequences were from metazoan organisms—except for a group of fungal sequences from ascomycetes. The resulting phylogenetic tree shows the dichotomy between the ZC3H4 and ZC3H6 domains, which are found in the same set of organisms. This indicates that they are paralogues and have likely diverged their function following gene duplication. The ancestral gene coding for ZC3H4/6 was likely lost from the non-vertebrates and subsequently underwent a duplication event leading to the ZC3H4- and ZC3H6-like paralogues in vertebrates. Primary data are available in Supplementary File 2 and deposited at Zenodo (https://doi.org/10.5281/zenodo.4637127).

b) Multiple sequence alignment of ZC3H4 and ZC3H6 homologs. PTHR13119 domains from human and mouse SwissProt sequences were aligned using structural information (PDB

structure: 2CQE; zinc-finger domain; helices are displayed as coils) using TCoffee (Expresso mode). Conserved regions are indicated by blue boxes; identical and similar residues (based on physicochemical properties) are marked in red and yellow, respectively. Sequence identifiers correspond to UniProt/SwissProt accession numbers and to boundaries of identified PTHR13119 domains. Alignment figure was rendered with ESPscript.

Figure 2 – figure supplement 1

a) qRT-PCR and western blotting evidence of the effectiveness of ZC3H6 and ZC3H4 depletion respectively. Graph shows fold reduction of ZC3H6 mRNA in cells treated with ZC3H6 siRNAs versus those transfected with control siRNA. N=3, error bars are SEM. ** is $p<0.01$. Western blotting of ZC3H4 in HCT116 cells treated with control siRNAs or ZC3H4-targeting siRNAs. The blot was probed with a ZC3H4 antibody revealing strong depletion versus the EXOSC10 loading control.

b) Pearson's correlation of siControl, siZC3H4, siZC3H6 and siZC3H4+6 of RNAseq BAM files performed by DEEPTOOLS.

c) IGV traces exemplifying two genomic regions with clear RNA accumulation following ZC3H4 depletion. While this is also seen following ZC3H4/ZC3H6 co-depletion, it is not evident following the depletion of ZC3H6 alone. This was generally seen, supporting the correlation analysis in b). y-axis scale is RPKM.

Figure 2 – figure supplement 2

a) Heatmaps showing the effects of CPSF30-mAID or ZC3H4 depletion on read-through beyond protein-coding genes. Coloured scale bar indicates the magnitude of effect (log2 scale). Read-through was scored as a ratio of reads upstream (500bp) and downstream (1kb) of the PAS. The lists associated with this heatmap are provided in Supplementary File 3.

b) Graph showing the number of protein-coding genes with read-through enhancements of greater than 0.5, 1, 2 or 3 on a log2 fold scale. This illustrates that the effects of CPSF30-mAID loss are both wider spread and larger than those associated with depletion of ZC3H4.

c) Venn diagram showing the number of genes bioinformatically scored as having increased read-through (log2 fold of 1 or more) following CPSF30-mAID or ZC3H4 loss and those that are common between the two conditions.

d) IGV tracks of individual genes (*MAGED1* and *DLG3)* to exemplify the lack of 3' termination defect following ZC3H4 depletion. As a control for *bone fide* read-through, the same tracks are shown in samples obtained from *CPSF30-mAID* cells treated or not (3hr) with auxin. Grey bar indicates a change in scale (left scale for upstream of PAS; right for downstream) so that comparatively weaker read-through signals can be visualised next to the gene body. y-axis scale is RPKM.

e) CPSF30 depletion shows little effect at super-enhancers and PROMPTs. IGV track view of the *MYC* super-enhancer and PROMPT in RNA-seq data obtained from *CPSF30-mAID* cells treated or not (3hr) with auxin. y-axis scale is RPKM.

f) Plot showing the density of PAS sequences (AWTAAA) in PROMPTs upregulated or unaffected by ZC3H4 loss. Y-axis plots number of PAS sequences/kb.

Figure 3 – figure supplement 1

a) IGV snapshots showing examples of protein-coding genes selectively upregulated by ZC3H4 (*PJVK* and *ENO3*) or Integrator (*TM7SF2* and *GFPT2*). Y-axis represents RPKM.

b) Venn diagram showing representing the number of protein-coding transcripts (determined by DESEQ2) that show increased levels in previously published 4sU labelling experiments performed on HeLa cells depleted of INTS11 or ZC3H4 (Austenaa et al., 2021; Lykke-Andersen et al., 2020). Gene lists are provided in Supplementary File 5. Notably, manual curation of this list revealed the presence of false positive hits, especially in the INTS11 data, due to DESEQ2 scoring interference of transcription from neighbouring genes as upregulation.

c) Venn diagram showing the number of PROMPTs showing upregulation (log2 fold of 1 or more) following ZC3H4 or INTS1 loss from HCT116 cells and those that are common between the two conditions.

d) Graph showing the number of PROMPTs enhanced by greater than 0.5, 1, 2 or 3 on a log2 fold scale following the loss of ZC3H4 or INTS1. This illustrates that the effects of ZC3H4 loss are both wider spread and larger than those associated with depletion of INTS1. The list of targets in each case is provided in Supplementary File 6.

Figure 4 – figure supplement 1

a) Pol II ChIP over ITPRID2 and MYC PROMPT regions. Graphs plot Pol II occupancy as a percentage of input at amplicons ~2, 4 and 8kb upstream of each gene. N=4. Error bars are

SEM. * denotes p of <0.05. The schematic illustrates the approximate location of each primer pair.

b) qRT-PCR analysis of extended ITPRID2 and MYC PROMPTs (-8kb amplicons) in chromatin-associated RNA isolated from *ZC3H4-DHFR* cells grown with or without (4hr) TMP. Y-axis displays fold change versus +TMP following normalisation to spliced actin levels. n=3. Error bars are SEM. * and ** denote p values of <0.05 and 0.01 respectively.

c) Western blot showing acute depletion of PNUTS (the nuclear targeting subunit of PP1 phosphatase), tagged with an auxin-inducible degron (PNUTS-AID). Blot shows extracts from unmodified HCT116 cells and *PNUTS-AID* cells which were treated or not with auxin (3hr). WDR82 is used as a loading control.

d) qRT-PCR of PROMPT and SE transcripts in *PNUTS-AID* cells treated or not with auxin (3 hr). Graph shows fold change by comparison with non-auxin treated cells following normalisation to spliced actin transcripts. N=3. Error bars are SEM. * and ** denote p values of <0.05 and 0.01 respectively.

e) qRT-PCR analysis of PROMPT and SE transcripts in HCT116 cells treated with control siRNAs or siRNAs against both SETD1A and B. Quantitation shows fold change versus cells transfected with control siRNAs following normalisation to spliced actin levels. Note that the PROMPT and SE targets that are increased following ZC3H4 loss show relatively little change following depletion of SETD1A and B. The success of the RNAi is indicated by the strong reduction of SETD1A and B mRNAs. N=3. Error bars show SEM. ** denotes p value of <0.01.

Figure 5 – figure supplement 1

a) XRNAX analysis of ZC3H4 RNA binding in cells. Samples show input and those isolated following UV treatment or not. Bands representing each protein are labelled accordingly. ZC3H4 is recovered in a UV-dependent manner indicating that it is directly bound to RNA in cells. The same is true of EXOSC10 that, as an exoribonuclease, acts as a positive control. TCF4 is a DNA binding transcription factor and acts as a negative control.

b) ZC3H4 only marks transcribed super-enhancers. IGV track view of ZC3H4 and Pol II occupancy at two different super-enhancers, *DLAGAP1* present in both HEPG2 and HCT116 (top tracks) and the MYC super-enhancer (bottom tracks) that is only present in HCT116 cells. HEPG2 and HCT116 super-enhancer annotation is under each track as blue bars and was obtained from dbSUPER. Y-axis shows CPM.

Figure 6 – figure supplement 1

a) qRT-PCR of *ZC3H4-DHFR* cells transfected with MS2-IRES-GFP before growth in the presence or, to deplete ZC3H4, absence of TMP (4 hr). Graph shows percentage of each amplicon following TMP removal relative to that found in the presence of TMP following normalisation to spliced actin transcripts. N=3. Error bars show SEM. * denotes $p<0.05$.

b) qRT-PCR of HCT116 cells transfected with IRES-GFP and either a control beta-globin plasmid (NTC) or ZC3H4-MS2. The graph shows the percentage of GFP RNA versus control following normalisation to spliced actin transcripts. N=3. Error bars show SEM.

SUPPLEMENTARY FILES

Supplementary File 1

Mass spectrometry data associated with the Pol II-miniTurbo experiment.

Supplementary File 2

Underpinning data for phylogenetic analyses.

Supplementary File 3

Log2 fold changes in read-through following CPSF30 or ZC3H4 depletion from HCT116 cells.

Supplementary File 4

List of mRNAs that are upregulated following ZC3H4 depletion (nuclear RNA-seq) or INTS1 depletion (chromatin-associated RNA-seq) in HCT116 cells.

Supplementary File 5

List of mRNAs that are upregulated following ZC3H4 depletion (4sU RNA-seq; (Austenaa et al., 2021)) or INTS11 depletion (4sU TT-seq; (Lykke-Andersen et al., 2020)) in HeLa cells. Genes in each set were manually checked for upregulation (TRUE) or possible artefacts – primarily transcription from surrounding region into a gene that is consequently scored as upregulated (FALSE).

Supplementary File 6

List of PROMPTs that are upregulated following ZC3H4 depletion (nuclear RNA-seq) or INTS1 depletion (chromatin-associated RNA-seq) in HCT116 cells.

Supplementary File 7

Table of oligonucleotide and DNA sequences used in this study.

Source data file 1

Values (average, SEM, p-value) of data underpinning graphs within the paper.

Source data file 2

Uncropped western blot images.

Figure 1

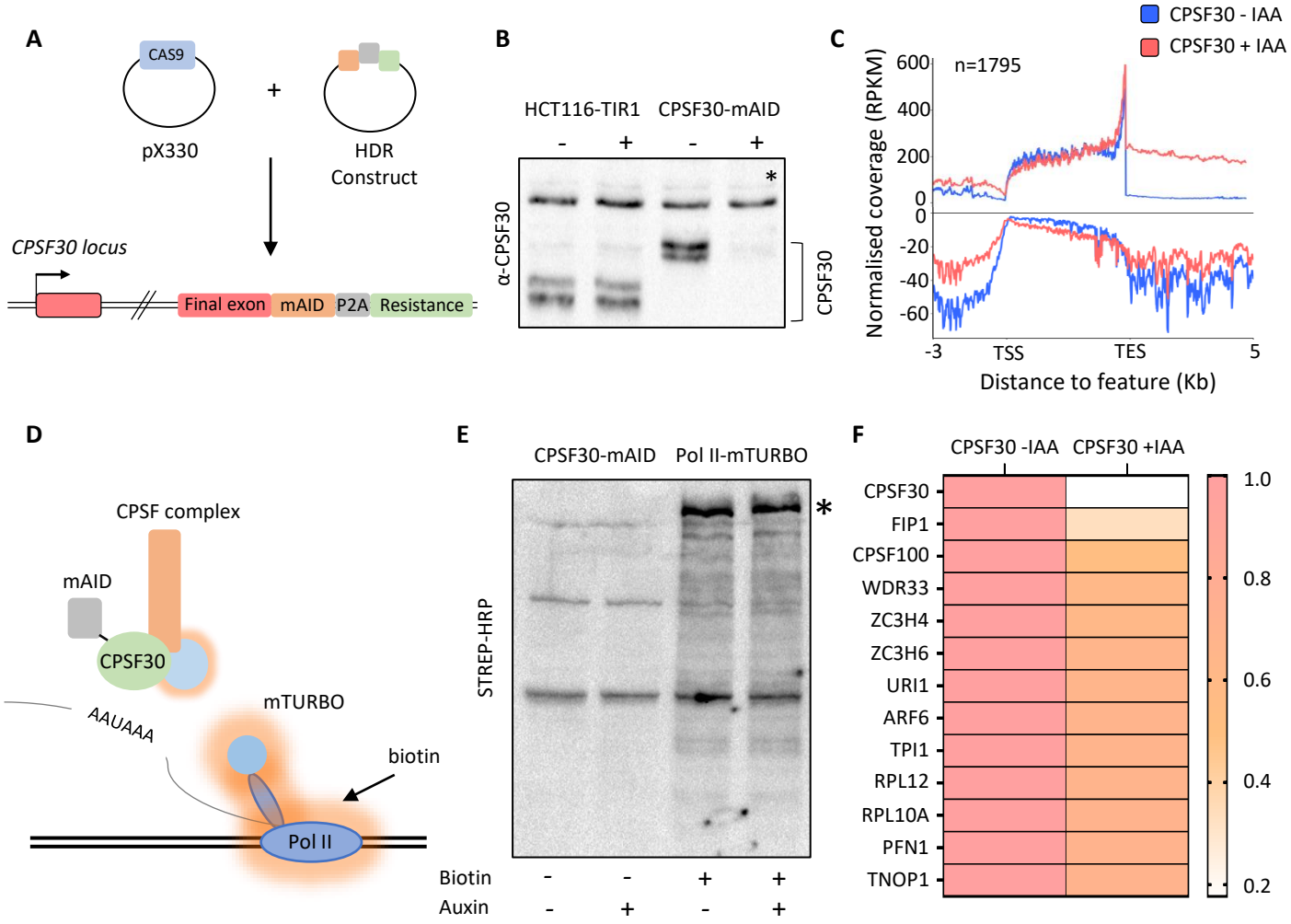
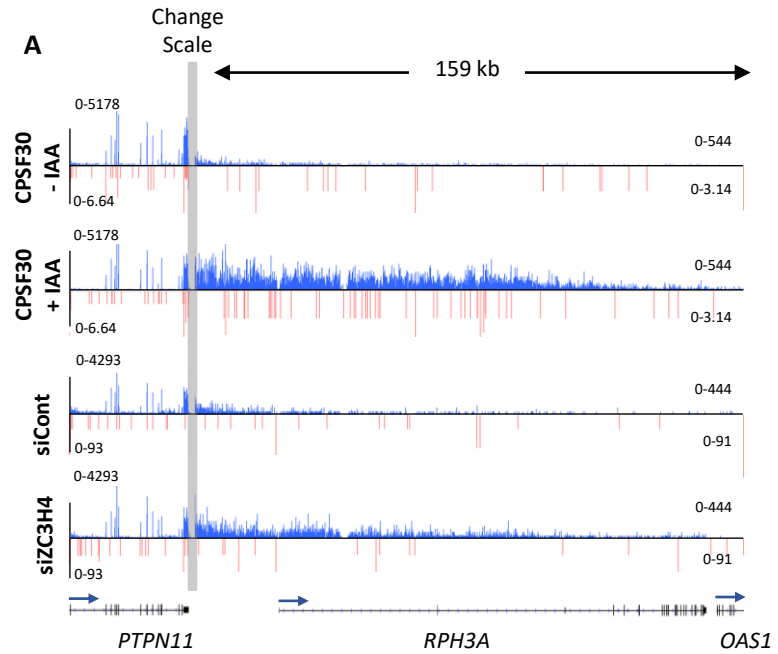
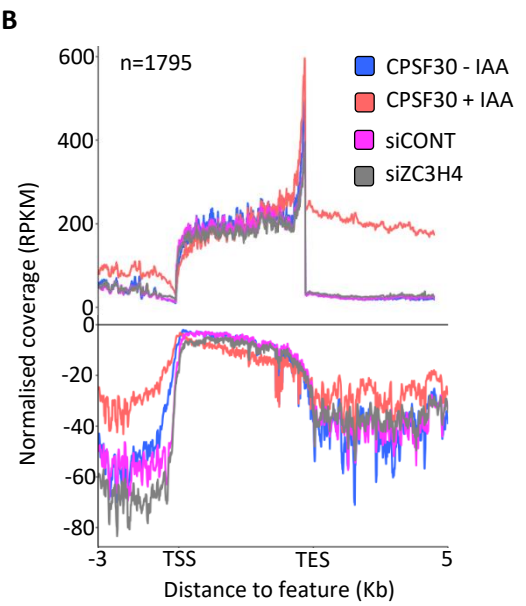
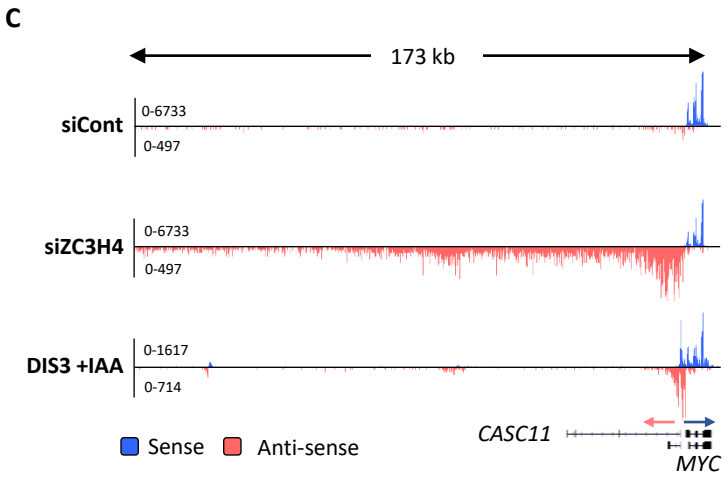
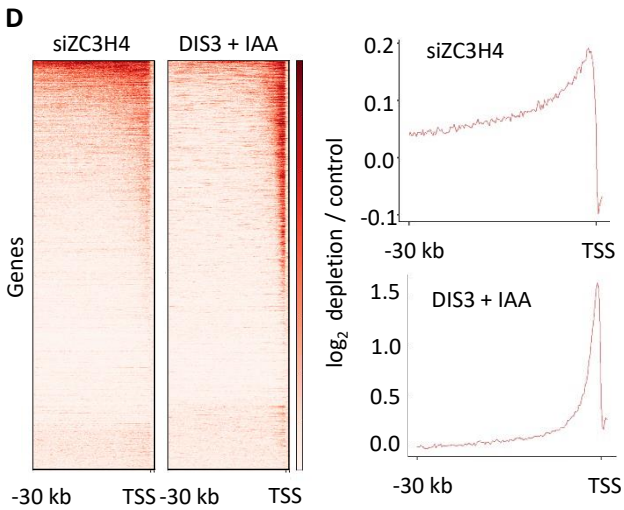

A
CAS9
pX330
+
HDR
Construct
CPSF30 locus
Final exon
mAID
P2A
Resistance
B
HCT116-TIR1
CPSF30-mAID
- + - +
α-CPSF30
*
CPSF30
C
CPSF30 - IAA
CPSF30 + IAA
n=1795
Normalised coverage (RPKM)
600
400
200
0
0
-20
-40
-60
-3
TSS
TES
5
Distance to feature (Kb)
D
CPSF complex
mAID
CPSF30
AAUAAA
mTURBO
biotin
Pol II
E
CPSF30-mAID
Pol II-mTURBO
STREP-HRP
*
Biotin - - + +
Auxin - + - +
F
CPSF30 -IAA
CPSF30 +IAA
CPSF30
FIP1
CPSF100
WDR33
ZC3H4
ZC3H6
URI1
ARF6
TPI1
RPL12
RPL10A
PFN1
TNOP1
1.0
0.8
0.6
0.4
0.2

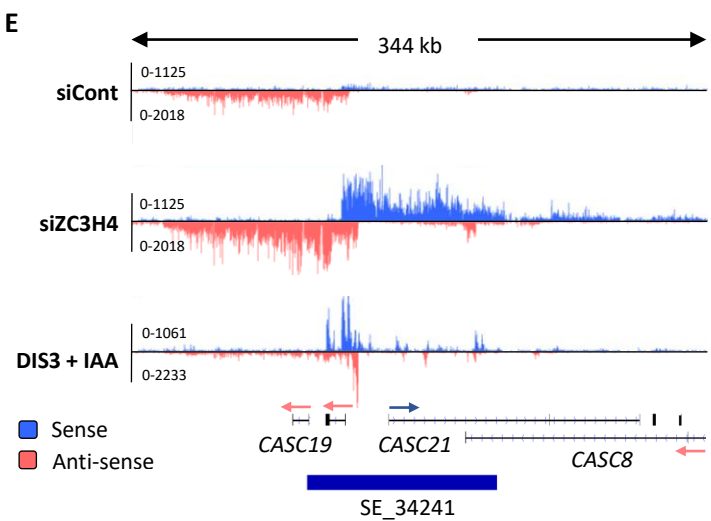
Figure 2


A
Change Scale
159 kb
CPSF30 - IAA
CPSF30 + IAA
siCont
siZC3H4
0-5178
0-6.64
0-544
0-3.14
0-4293
0-93
0-444
0-91
PTPN11
RPH3A
OAS1


B
n=1795
CPSF30 - IAA
CPSF30 + IAA
siCONT
siZC3H4
Normalised coverage (RPKM)
600
400
200
0
-20
-40
-60
-80
-3
TSS
TES
5
Distance to feature (Kb)


C
173 kb
siCont
siZC3H4
DIS3 +IAA
0-6733
0-497
0-1617
0-714
Sense
Anti-sense
CASC11
MYC


D
siZC3H4
DIS3 + IAA
Genes
-30 kb
TSS
log2 depletion / control
0.2
0.1
0.0
-0.1
1.5
1.0
0.5
0.0


E
344 kb
siCont
siZC3H4
DIS3 + IAA
0-1125
0-2018
0-1061
0-2233
Sense
Anti-sense
CASC19
CASC21
CASC8
SE_34241

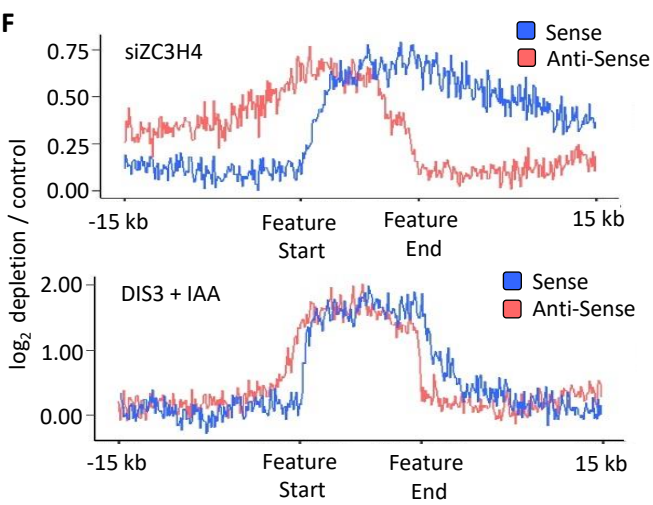

F
siZC3H4
DIS3 + IAA
Sense
Anti-Sense
log2 depletion / control
0.75
0.50
0.25
0.00
2.00
1.00
0.00
-15 kb
Feature Start
Feature End
15 kb

Figure 3

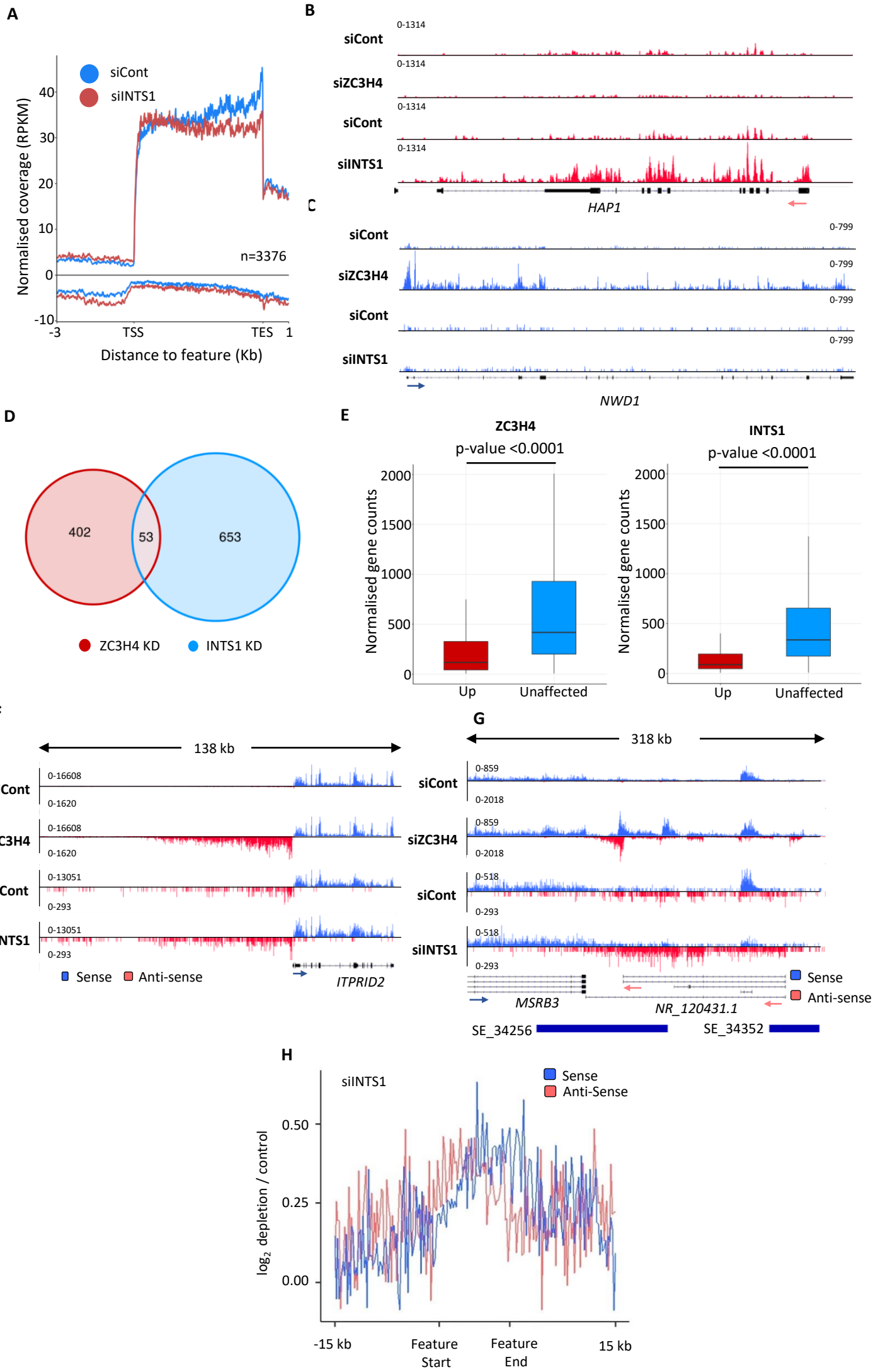

A
siCont
siINTS1
Normalised coverage (RPKM)
40
30
20
10
0
-10
n=3376
-3
TSS
TES
1
Distance to feature (Kb)
B
siCont
siZC3H4
siCont
siINTS1
0-1314
HAP1
C
siCont
siZC3H4
siCont
siINTS1
0-799
NWD1
D
402
53
653
ZC3H4 KD
INTS1 KD
E
ZC3H4
p-value <0.0001
INTS1
p-value <0.0001
Normalised gene counts
2000
1500
1000
500
0
Up
Unaffected
F
138 kb
siCont
siZC3H4
siCont
siINTS1
0-16608
0-1620
0-13051
0-293
Sense
Anti-sense
ITPRID2
G
318 kb
0-859
0-2018
0-518
0-293
MSRB3
NR_120431.1
SE_34256
SE_34352
H
siINTS1
Sense
Anti-Sense
log2 depletion / control
0.50
0.25
0.00
-15 kb
Feature Start
Feature End
15 kb

Figure 4

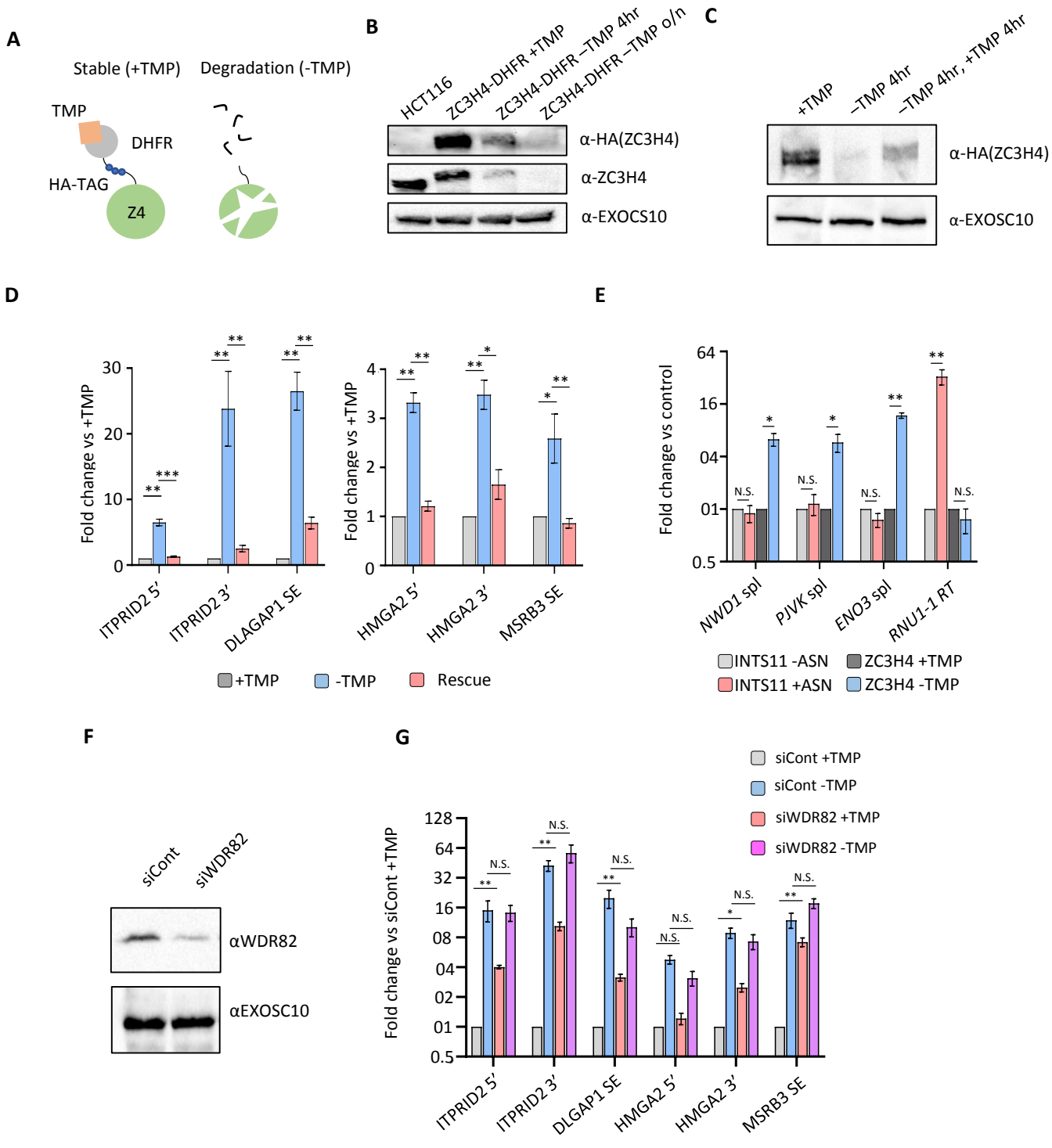

A
Stable (+TMP)
Degradation (-TMP)
TMP
DHFR
HA-TAG
Z4
B
HCT116
ZC3H4-DHFR +TMP
ZC3H4-DHFR –TMP 4hr
ZC3H4-DHFR –TMP o/n
α-HA(ZC3H4)
α-ZC3H4
α-EXOCS10
C
+TMP
–TMP 4hr
–TMP 4hr, +TMP 4hr
α-HA(ZC3H4)
α-EXOSC10
D
Fold change vs +TMP
ITPRID2 5'
ITPRID2 3'
DLAGAP1 SE
HMGA2 5'
HMGA2 3'
MSRB3 SE
+TMP
-TMP
Rescue
E
Fold change vs control
NWD1 spl
PJVK spl
ENO3 spl
RNU1-1 RT
INTS11 -ASN
INTS11 +ASN
ZC3H4 +TMP
ZC3H4 -TMP
F
siCont
siWDR82
αWDR82
αEXOSC10
G
Fold change vs siCont +TMP
siCont +TMP
siCont -TMP
siWDR82 +TMP
siWDR82 -TMP
ITPRID2 5'
ITPRID2 3'
DLGAP1 SE
HMGA2 5'
HMGA2 3'
MSRB3 SE

Figure 5

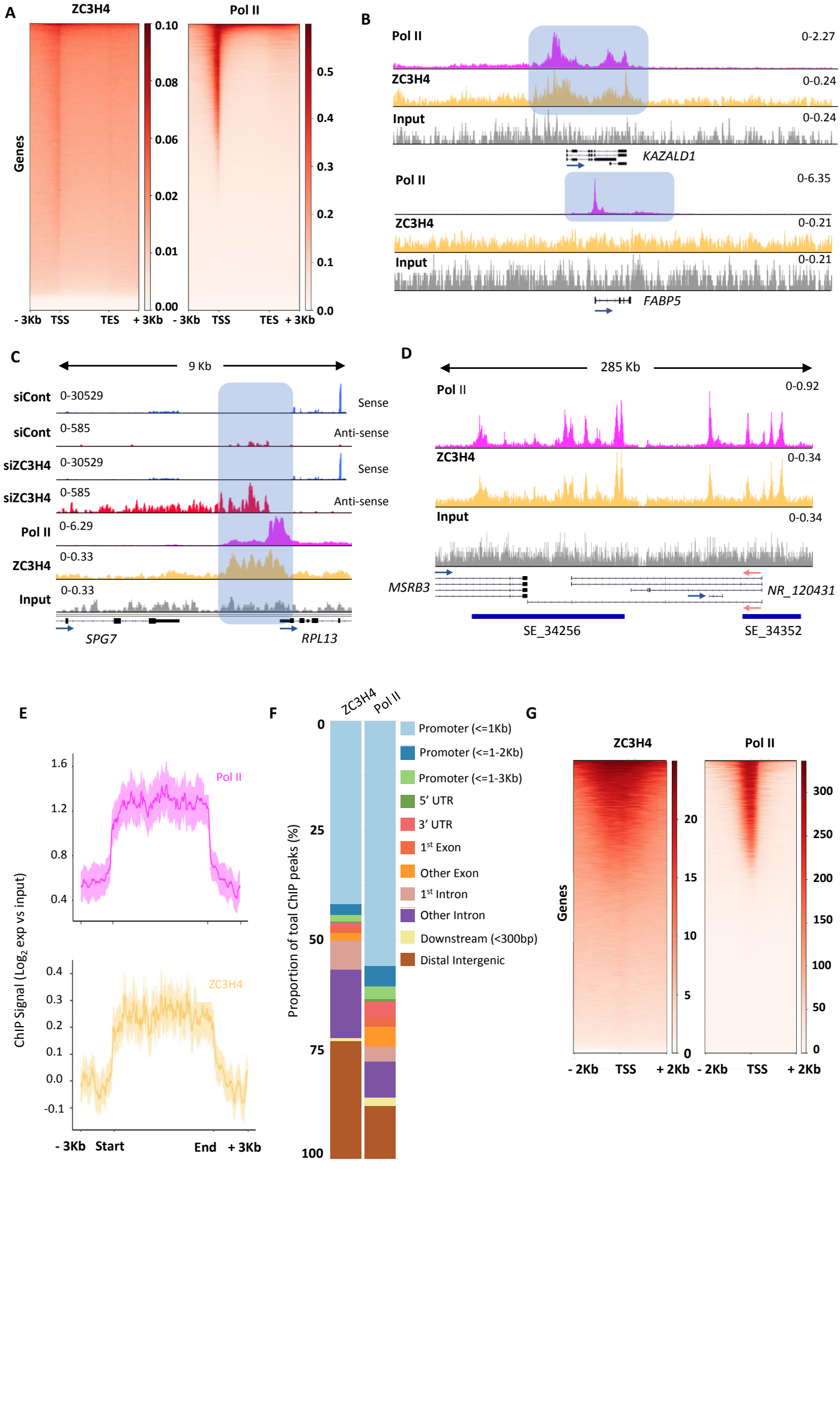

A
ZC3H4
Pol II
Genes
0.10
0.08
0.06
0.02
0.01
0.00
0.5
0.4
0.3
0.2
0.1
0.0
- 3Kb TSS TES + 3Kb
- 3Kb TSS TES + 3Kb
B
Pol II
0-2.27
ZC3H4
0-0.24
Input
0-0.24
KAZALD1
Pol II
0-6.35
ZC3H4
0-0.21
Input
0-0.21
FABP5
C
9 Kb
siCont
0-30529
Sense
siCont
0-585
Anti-sense
siZC3H4
0-30529
Sense
siZC3H4
0-585
Anti-sense
Pol II
0-6.29
ZC3H4
0-0.33
Input
0-0.33
SPG7
RPL13
D
285 Kb
Pol II
0-0.92
ZC3H4
0-0.34
Input
0-0.34
MSRB3
NR_120431
SE_34256
SE_34352
E
Pol II
ZC3H4
ChIP Signal (Log2 exp vs input)
1.6
1.2
0.8
0.4
0.4
0.3
0.2
0.1
0.0
-0.1
- 3Kb Start
End + 3Kb
F
ZC3H4
Pol II
Proportion of toal ChIP peaks (%)
0
25
50
75
100
Promoter (<=1Kb)
Promoter (<=1-2Kb)
Promoter (<=1-3Kb)
5′ UTR
3′ UTR
1st Exon
Other Exon
1st Intron
Other Intron
Downstream (<300bp)
Distal Intergenic
G
ZC3H4
Pol II
Genes
20
15
10
5
0
300
250
200
150
100
50
0
- 2Kb TSS + 2Kb
- 2Kb TSS + 2Kb

Figure 6

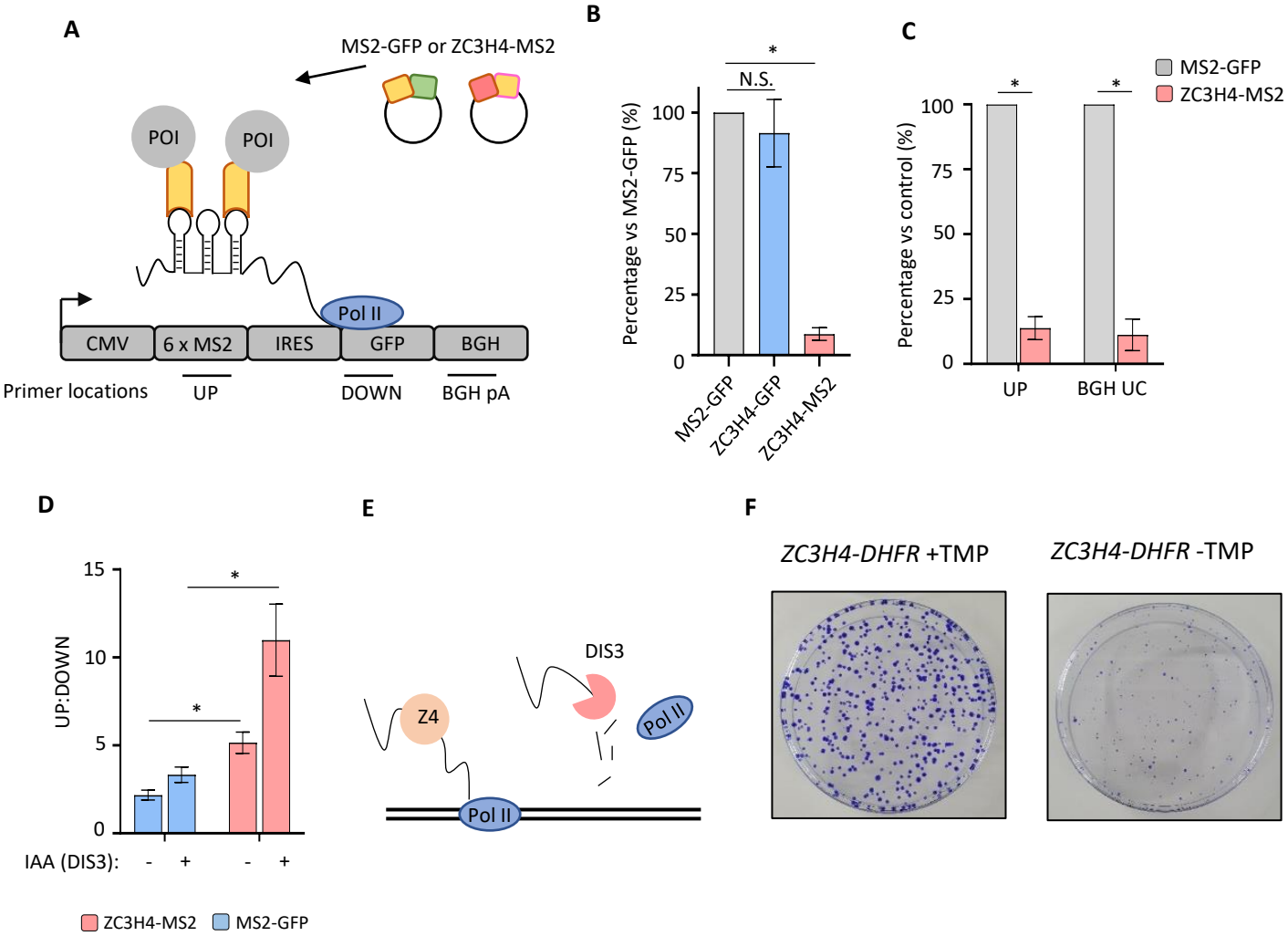
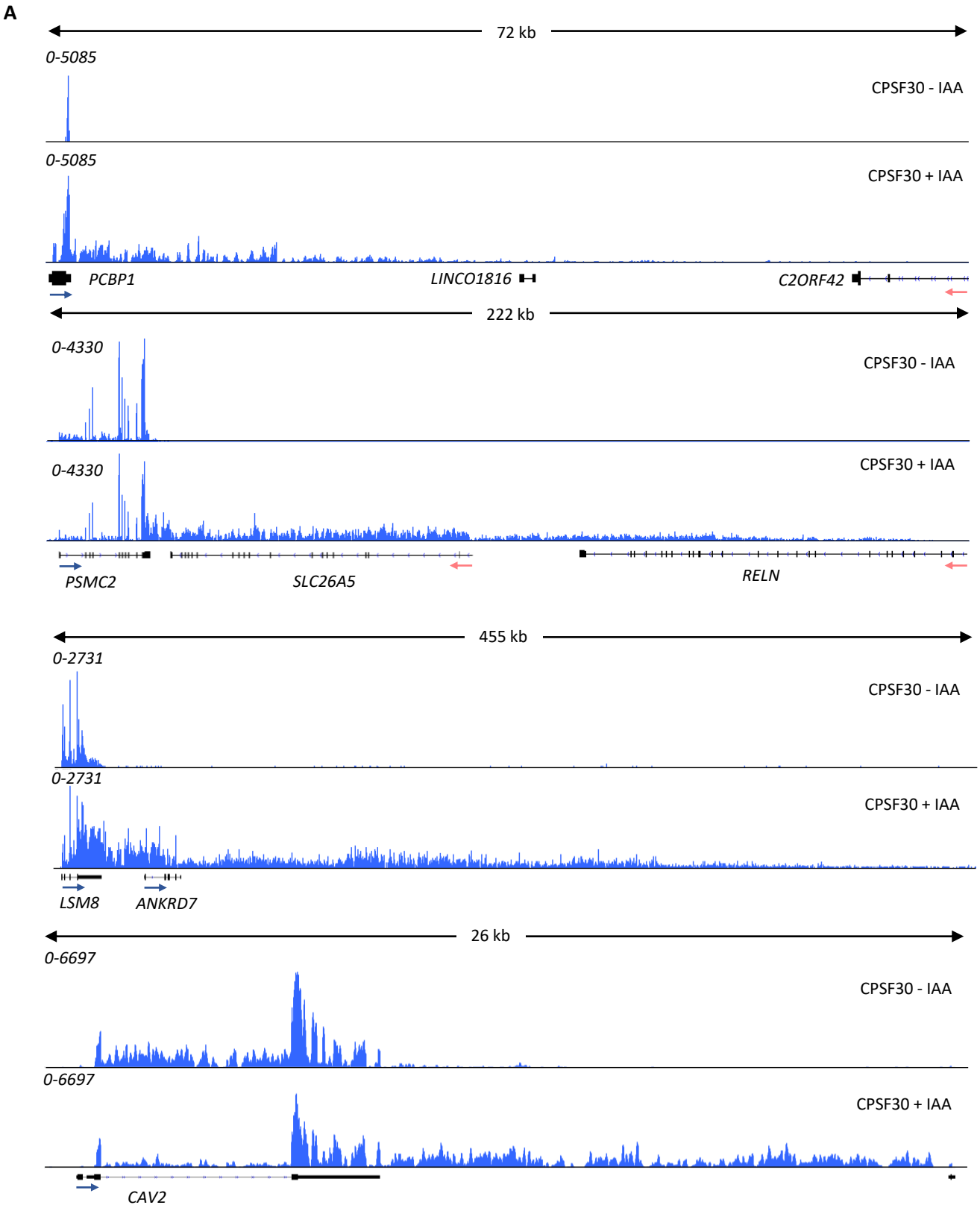
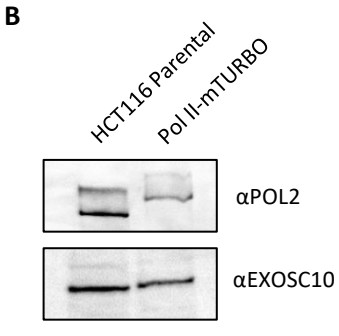

A
MS2-GFP or ZC3H4-MS2
POI
POI
Pol II
CMV
6 x MS2
IRES
GFP
BGH
Primer locations
UP
DOWN
BGH pA
B
*
N.S.
Percentage vs MS2-GFP (%)
100
75
50
25
0
MS2-GFP
ZC3H4-GFP
ZC3H4-MS2
C
MS2-GFP
ZC3H4-MS2
*
*
Percentage vs control (%)
100
75
50
25
0
UP
BGH UC
D
*
*
UP:DOWN
15
10
5
0
IAA (DIS3):
-
+
-
+
ZC3H4-MS2
MS2-GFP
E
DIS3
Z4
Pol II
Pol II
F
ZC3H4-DHFR +TMP
ZC3H4-DHFR -TMP

Figure 1-figure supplement 1





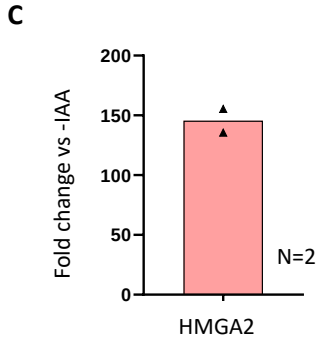
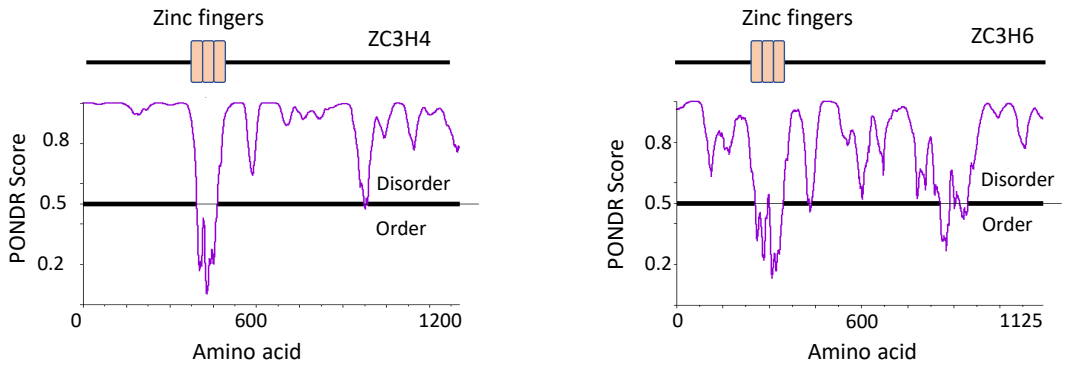
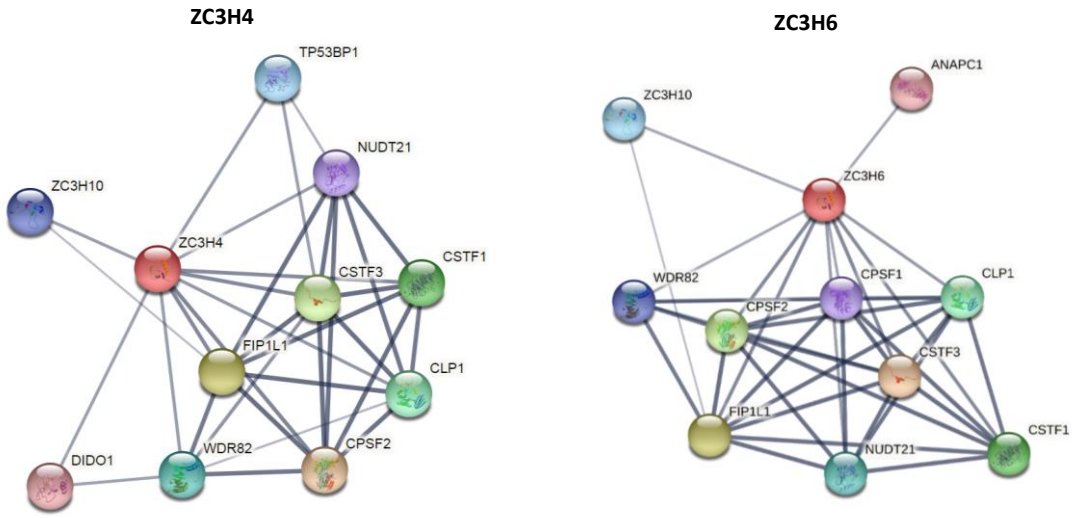


Figure 1-figure supplement 2

A



B



C

GO Biological Processes enriched among 19 identified proteins	No. of proteins	Adj p-value
RNA metabolic process	16	4.46E-06
gene expression	15	4.42E-04
mRNA processing	7	6.62E-04
cellular macromolecule metabolic process	17	4.58E-03
RNA splicing	6	6.44E-03
RNA biosynthetic process	11	9.45E-02

D

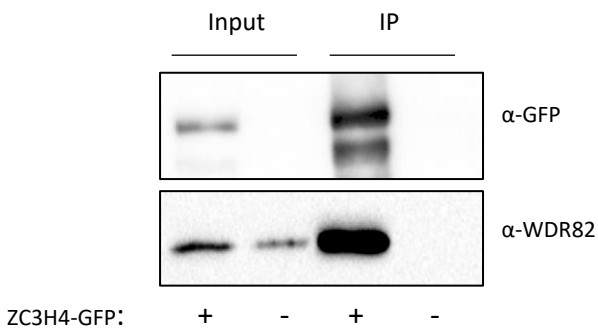
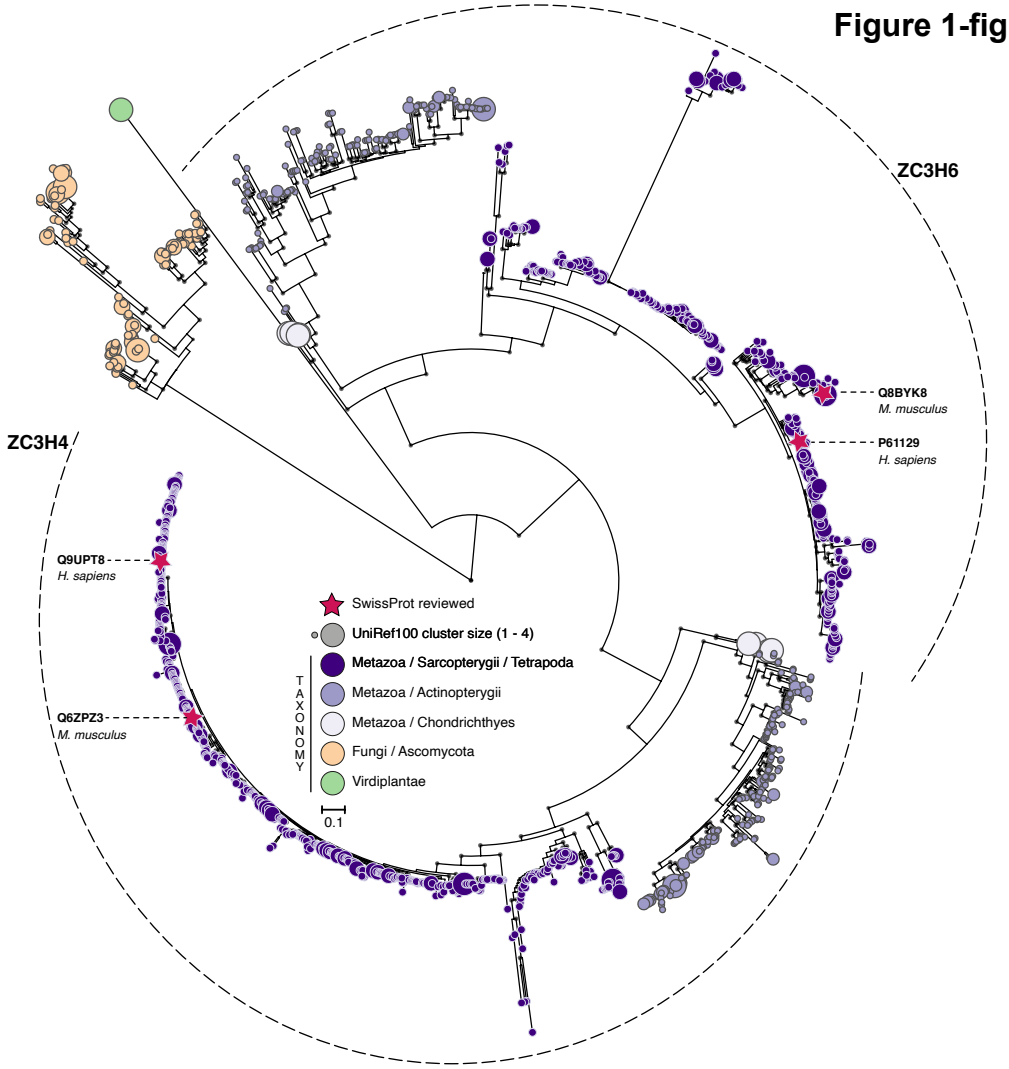


Figure 1-figure supplement 3

A



B

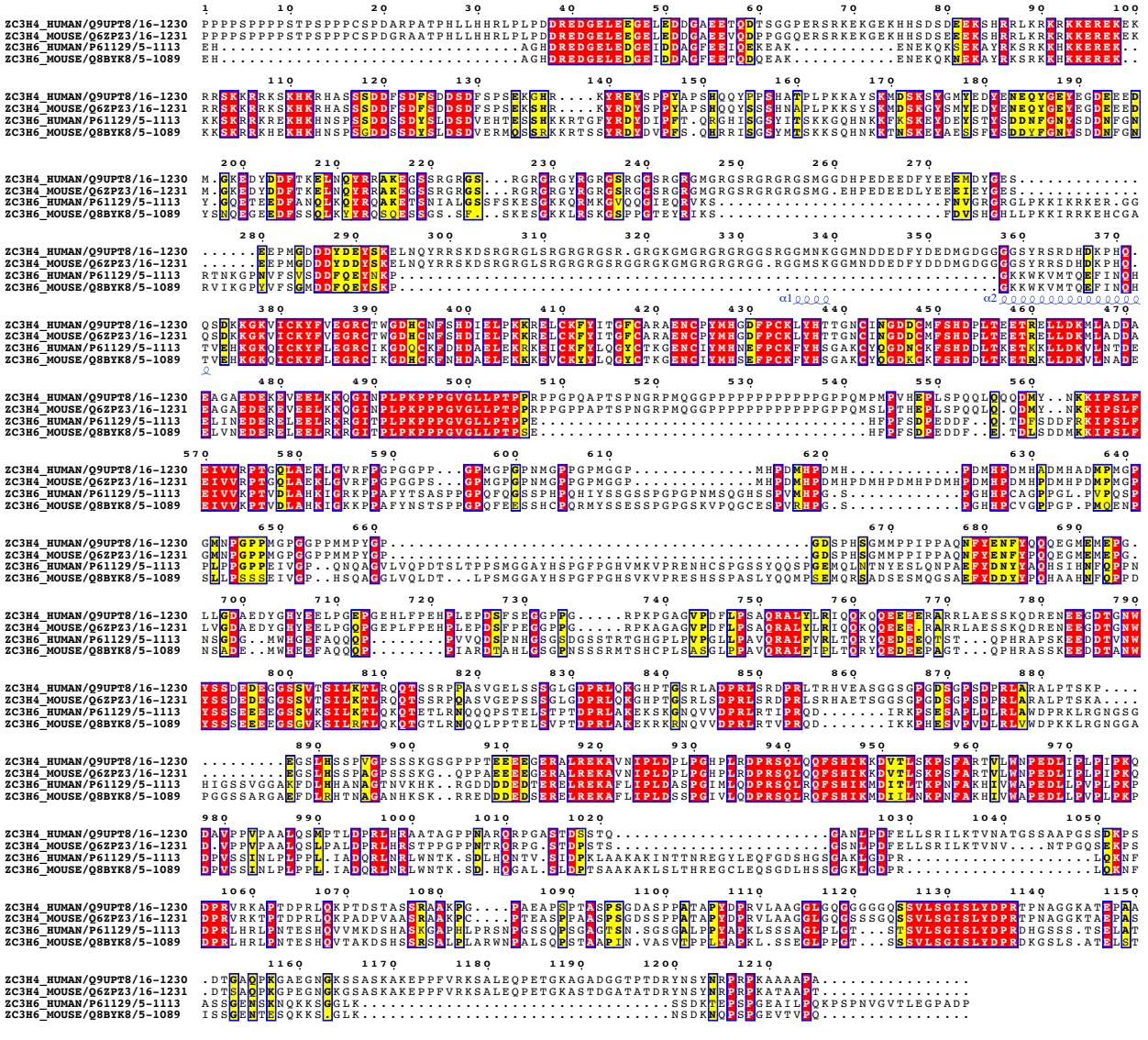


Figure 2-figure supplement 1

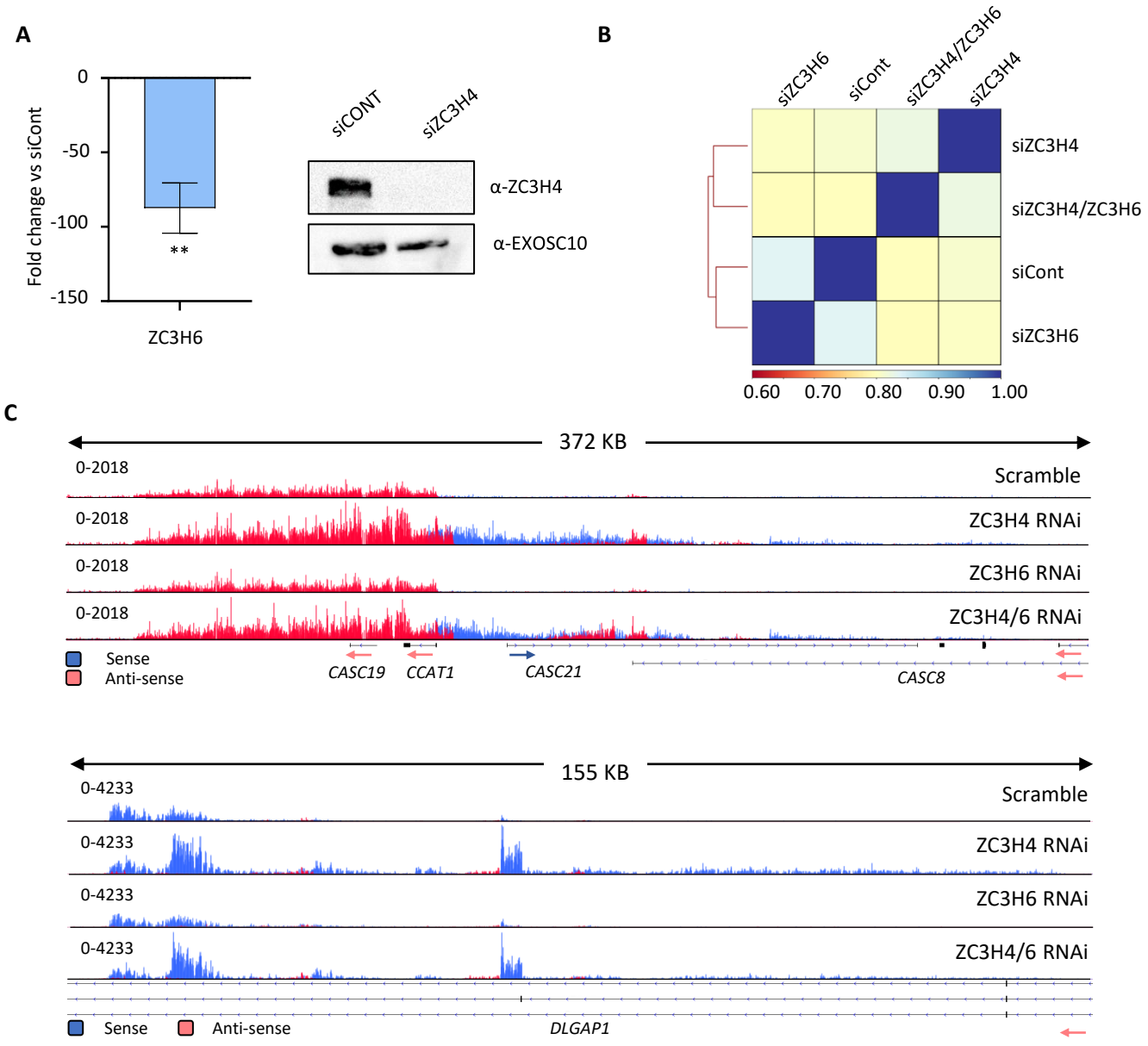


Figure 2-figure supplement 2

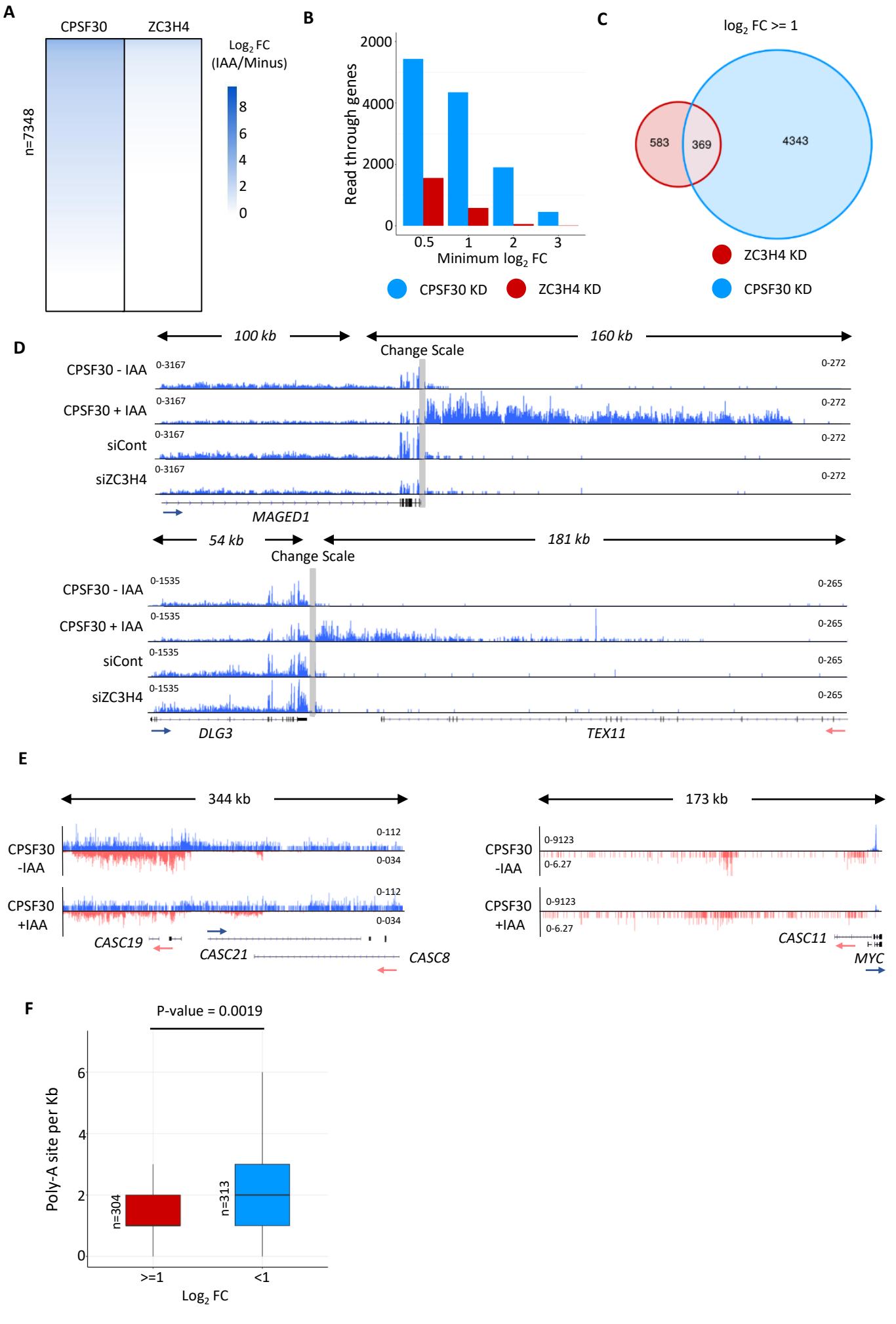


Figure 3-figure supplement 1

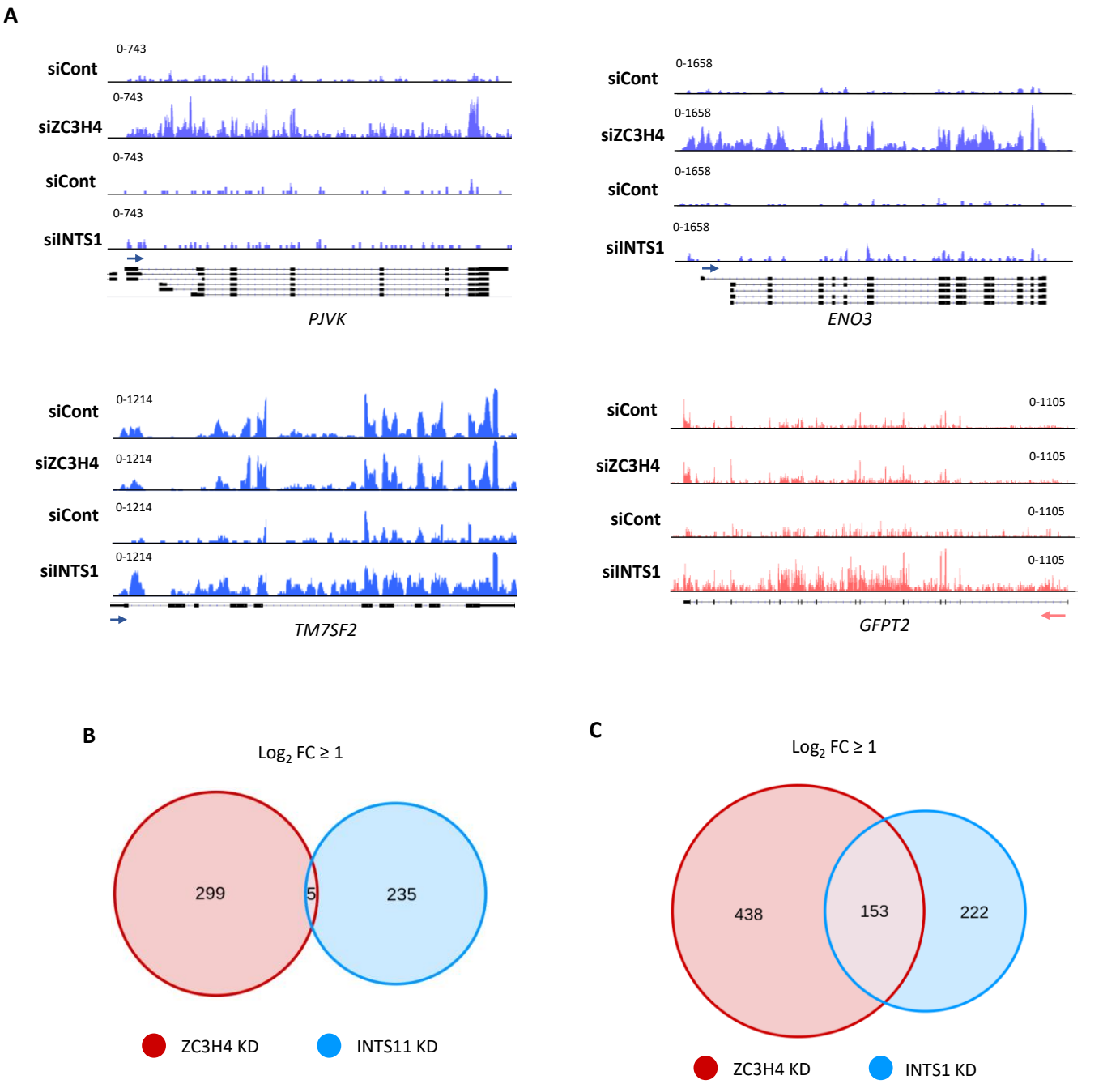


Figure 4-figure supplement 1

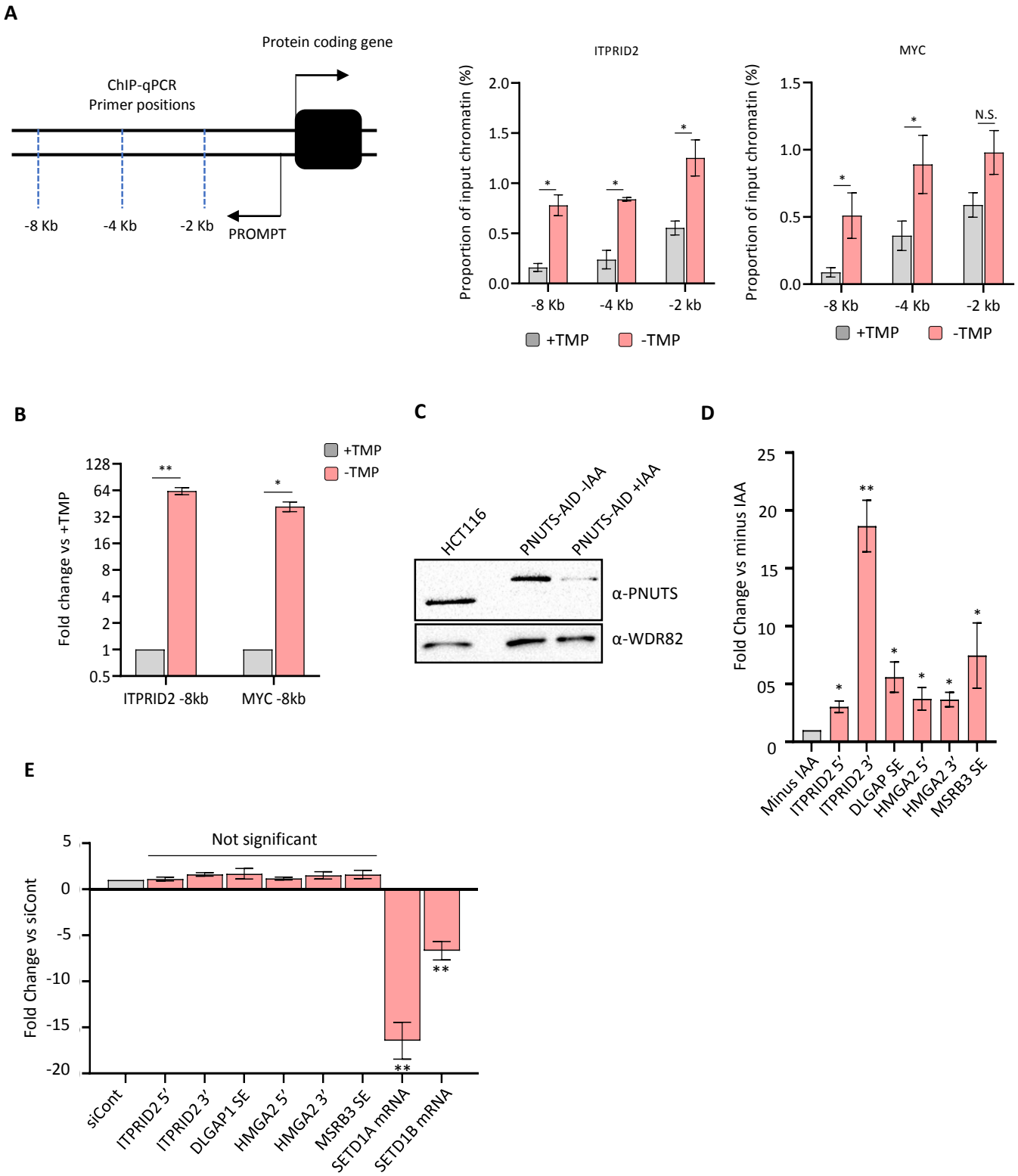


Figure 5-figure supplement 1

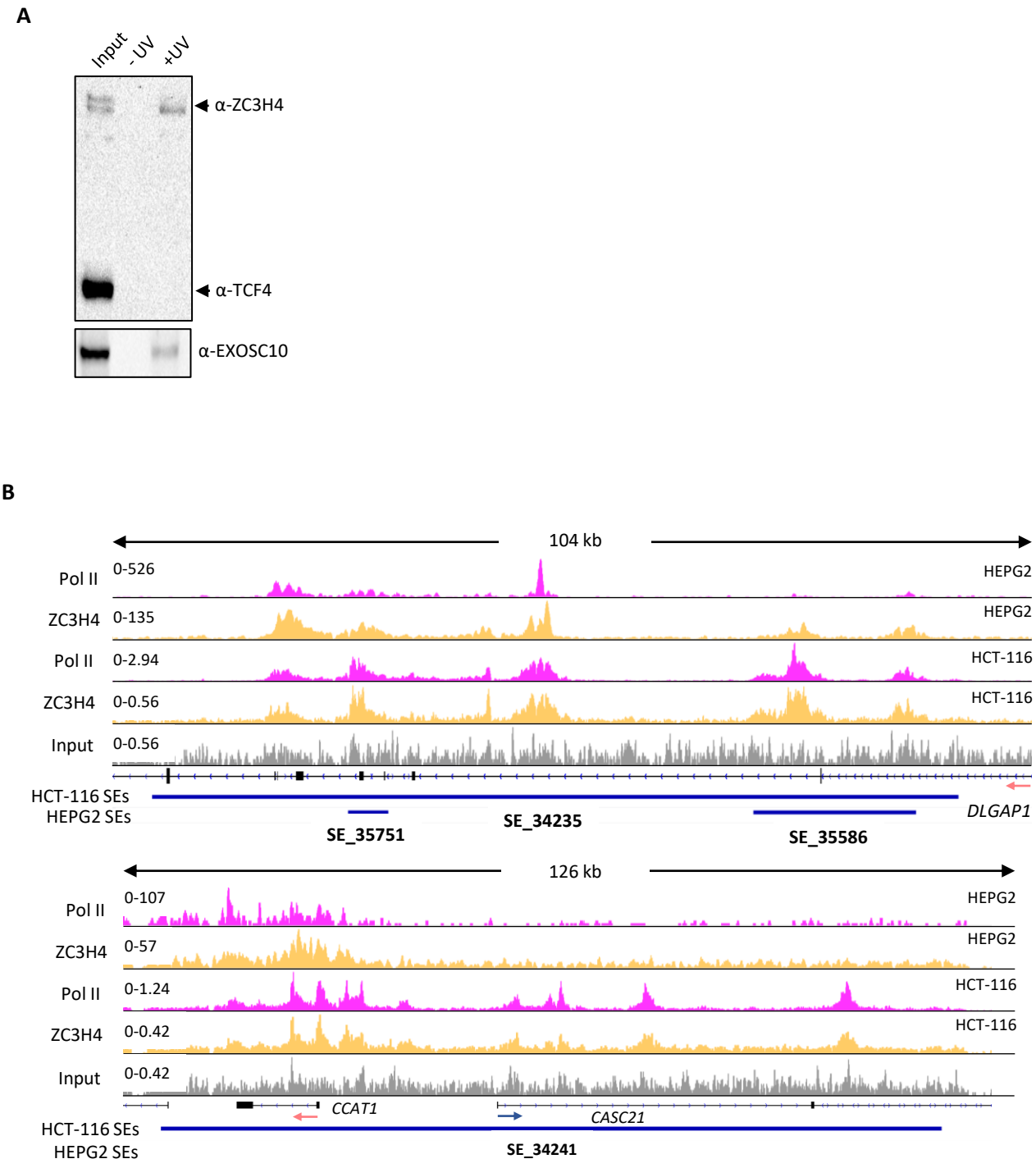


Figure 6-figure supplement 1